Silvano Salvador

# Toponomastic Networks

*A Graph-Theoretic Framework for the Topology of Phonotactic Substrate Relations in Toponomastic Corpora*

Methodological Volume — Graph-Theoretic Foundations

*A Companion Series to Toponomastic Entropy*

# Contents

**Preface** **xi**

**1 From Pairwise Divergence to Network Structure** **1**
1.1 The Limits of the Pairwise Framework . . . . . 1
1.2 What a Graph Adds . . . . . 2
1.3 Outline of the Volume . . . . . 2

**2 Network Science in Linguistics: A Literature Review** **4**
2.1 Word Co-occurrence and Semantic Networks . . . . . 4
2.2 Phonological and Phonotactic Networks . . . . . 5
2.3 Dialectometry and Areal Linguistics . . . . . 5
2.4 Language Contact and Areal Diffusion as Network Phenomena . . . 6
2.5 What Is Genuinely New in This Volume's Approach . . . . . 6

**3 The Metric Structure of the Edge-Weight Function** **7**
3.1 Metric Axioms . . . . . 7
3.2 Jensen–Shannon Divergence Is a True Metric . . . . . 8
3.3 Why Triangle-Inequality Violations Would (and Would Not) Matter 9

**4 Graph-Theoretic Foundations** **10**
4.1 Weighted Graphs . . . . . 10
4.2 Degree and Local Structure . . . . . 11
4.3 Paths, Geodesics, and Betweenness Centrality . . . . . 11
4.4 Eigenvector Centrality . . . . . 12
4.5 Community Detection: Modularity and the Louvain Algorithm . . . 13
4.6 Spectral Graph Theory: The Graph Laplacian . . . . . 14
4.7 Minimum Spanning Trees . . . . . 14

**5 Extended Graph-Theoretic Toolkit** **15**
5.1 Local Clustering: The Clustering Coefficient . . . . . 15
5.2 Distance Structure: Diameter, Radius, and Average Path Length . . . 16
5.3 Degree Assortativity . . . . . 17
5.4 Random Graph Null Models . . . . . 18
5.5 Robustness and Percolation . . . . . 19
5.6 The Small-World Coefficient . . . . . 20
5.7 Summary: Which Tools Apply Where . . . . . 20

**6 The Special Case of Complete Weighted Graphs** **21**
6.1 What Completeness Removes from the Analysis . . . . . 21
6.2 Weighted Degree (Strength) as the Natural Replacement . . . . . 22
6.3 The Complete Graph's Laplacian Spectrum in Closed Form . . . . . 23
6.4 Why a Sparsifying Step Is Necessary for Several of This Volume's Tools 23

**7 Constructing the Thirty-Three-Zone Graph** **25**
7.1 The Vertex Set . . . . . 25
7.2 Intra-Regional Edges: Inherited Real KL Divergences . . . . . 25
7.3 Inter-Regional Edges: A Declared Proxy Distance . . . . . 26
7.4 Basic Descriptive Statistics of the Full Graph . . . . . 27
7.5 Sensitivity of Construction Parameters . . . . . 27

**8 Alternative Centrality Measures: Theoretical Foundations** **31**
8.1 Closeness Centrality . . . . . 31
8.2 Katz Centrality and PageRank . . . . . 32
8.3 Subgraph Centrality and Walk-Counting Measures . . . . . 33
8.4 A Taxonomy of Centrality by What Each Measure Counts . . . . . 34
8.5 The General Principle: No Centrality Measure Is Privileged . . . . . 34

**9 Centrality: Which Zones Are Structurally Central?** **35**
9.1 Three Notions of Centrality, Computed . . . . . 35
9.2 Agreement and Disagreement Between Measures . . . . . 36
9.3 The Most Peripheral Zones . . . . . 36

**10 Community Detection: Theoretical Foundations Extended** **38**
10.1 The Stochastic Block Model . . . . . 38
10.2 Validating Detected Communities: Normalised Mutual Information . 39
10.3 Alternative Algorithms: Infomap and Label Propagation . . . . . 40
10.4 Modularity's Statistical Significance . . . . . 41

**11 Community Detection: Does the Network Know Geography?** **43**
11.1 Louvain Communities on the Intra-Regional Graph . . . . . 43
11.2 Louvain Communities on the Full Graph . . . . . 43
11.3 The Fifteen-Zone Merged Community: A Structural Finding . . . . . 44
11.4 Comparison to the Regional Partition: A Quantitative Summary . . . 45

**12 A Declared Geographic Contiguity Graph** **47**
12.1 Why a Declared Construction, and What It Is Not . . . . . 47
12.2 The Edge List, by Region . . . . . 48
12.3 Basic Structure of $\mathcal{G}_{\text{geo}}$ . . . . . 49
12.4 Comparing $\mathcal{G}_{\text{geo}}$ to the Phonotactic Graph's Partitions . . . . . 49
12.5 Agenda Item: From Declared Skeleton to Measured Polygon Graph . 51

**13 Structural Diagnostics on the Declared Geographic Graph** **52**
13.1 Degree Assortativity of $\mathcal{G}_{\text{geo}}$ . . . . . 52
13.2 Clustering and Cycles . . . . . 52
13.3 Why a Global Small-World Coefficient Is Not Reported . . . . . 53

**14 Spectral Graph Theory in Depth** **55**
14.1 The Normalised Laplacian . . . 55
14.2 Cheeger's Inequality . . . 56
14.3 Eigenvalue Interlacing . . . 57
14.4 The Matrix-Tree Theorem . . . 57

**15 Global Topology: Backbone, Spectrum, and Robustness** **59**
15.1 The Minimum Spanning Tree . . . 59
15.2 A Formalized Cross-Volume Dendrogram . . . 60
15.3 Spectral Connectivity: The Fiedler Value . . . 62
15.4 Sensitivity Analysis: Robustness to the Proxy Parameters . . . 63

**16 Algorithmic Considerations and Computational Complexity** **65**
16.1 Shortest Paths and Betweenness Centrality . . . 65
16.2 Eigenvector Centrality via Power Iteration . . . 66
16.3 Louvain Community Detection . . . 67
16.4 Minimum Spanning Tree: Kruskal's Algorithm and Union-Find . . . 68
16.5 Spectral Computations: Eigendecomposition Cost . . . 68
16.6 A Summary Table of Complexities . . . 68

**17 Synthesis and Agenda for the Series** **70**
17.1 What the Graph-Theoretic Reading Adds to the Parent Series . . . 70
17.2 The Central Limitation, Restated . . . 70
17.3 Agenda for Volumes II–IV . . . 71

**18 Full Proofs I: Centrality and Walk-Counting Measures** **73**
18.1 Perron–Frobenius and the Existence of Eigenvector Centrality . . . 73
18.2 Convergence of Power Iteration . . . 74
18.3 Katz Centrality as a Resolvent Series . . . 75
18.4 PageRank as a Perturbed Stochastic Eigenproblem . . . 76
18.5 Subgraph Centrality and the Matrix Exponential . . . 76

**19 Full Proofs II: Spectral Graph Theory** **78**
19.1 Cheeger's Inequality: Both Directions . . . 78
19.2 Cauchy Interlacing in Full . . . 79
19.3 The Matrix-Tree Theorem in Full . . . 80

**20 Algorithms in Full: Pseudocode and Complexity Derivations** **81**
20.1 Dijkstra's Algorithm and Brandes' Betweenness Algorithm . . . 81
20.2 Power Iteration for Eigenvector Centrality . . . 82
20.3 Louvain Community Detection . . . 82
20.4 Kruskal's Algorithm with Union–Find . . . 82
20.5 Eigendecomposition Cost for Spectral Methods . . . 84

**21 Independent Recomputation from Raw GeoNames Corpora** **85**
21.1 Extraction Pipeline and a Necessary Methodological Caveat . . . 85
21.2 Validation Against the Inherited Intra-Regional Values . . . 86
21.3 A Fully Data-Driven Graph: Real Inter-Regional Edges . . . 87
21.4 Centrality and Community Structure, Recomputed . . . 88

21.5 Minimum Spanning Tree and Spectral Connectivity, Recomputed . . 89
21.6 Percolation and Robustness: Real Numbers at Last . . . . . 89
21.7 Synthesis: What Changes, What Is Confirmed, What Remains Open . 91

**22 A Country-Level Cross-Check: Coarsening the Vertex Set 92**
22.1 Extraction Pipeline . . . . . 92
22.2 The Pairwise Divergence Matrix . . . . . 93
22.3 Centrality on the Country-Level Graph . . . . . 94
22.4 Community Structure: Does the Country-Level Partition Survive? . . 95
22.5 Minimum Spanning Tree, Spectral Connectivity, and Two New Graph Statistics . . . . . 96
22.6 Percolation, Recomputed a Second Time . . . . . 98
22.7 Synthesis: What the Coarsening Confirms, Changes, and Cannot Settle 99

**23 Geography Versus Phonotactics: A Direct Correlation Test 102**
23.1 A Real Geographic Distance Matrix from Corpus Centroids . . . . . 102
23.2 Mantel Correlation Test . . . . . 102

**24 Sensitivity to the Smoothing Parameter $\alpha$ 105**

**25 Bigram Vocabulary Overlap: A Type-Based Complementary Measure 107**
25.1 Computation and Basic Statistics . . . . . 107
25.2 Correlation with Token-Frequency Divergence . . . . . 108

**26 Full Proofs III: Properties of the New Chapter-22–25 Statistics 110**
26.1 Bounds on the Jaccard Index and Its Relation to Divergence . . . . . 110
26.2 Bounds on the Weighted Assortativity Coefficient . . . . . 112
26.3 The Weighted Clustering Coefficient: Range and a Degenerate Case . 112
26.4 The Mantel Test: Why Permutation Rather Than a Closed-Form $p$-Value 113

**27 Appendix: Reproducibility — Pipeline Pseudocode for Chapters 22–25 114**

**28 Integrated Synthesis: Four Independent Lines of Evidence 117**
28.1 The Single Most Robust Finding in This Volume: Libya's Isolation . . 117
28.2 A Secondary, Two-Way-Corroborated Finding: Iberian/Maghreb Centrality . . . . . 118
28.3 The Genuine Null Result: Geography Does Not Predict Phonotactics Globally . . . . . 119
28.4 What Remains for a Future Volume . . . . . 119

**Bibliography 120**

**Appendix: The Full Country-Level Pairwise Table 121**

**Appendix: Distinctive Bigram Signatures by Country 127**
28.5 Method: A Log-Likelihood Distinctiveness Score . . . . . 127
28.6 Reading the Signatures . . . . . 128

**Appendix: Sensitivity of the Country-Level Results to Sample Size $n$** **130**
28.7 Entropy as a Function of Sample Size . . . . . 130
28.8 Stability of the Divergence Matrix and Downstream Findings . . . . . 131
28.9 The Null Mantel Result Is Also Stable Across $n$ . . . . . 133

**Appendix: Notation Index** **134**

# List of Tables

5.1 Where each tool of this chapter is used later in the volume. . . . . 20

7.1 The thirty-three vertices of the graph $\mathcal{G}$, grouped by regional volume of origin, with bias-corrected Shannon bigram entropy $H_2$ (bits), inherited and verified from *Toponomastic Entropy*, Vol. I, Table 11.1. . . . 29

7.2 Summary of the six intra-regional symmetrised KL divergence matrices inherited from *Toponomastic Entropy*, Vols. II–VII. $\bar{\delta}$: mean symmetrised divergence; range: minimum–maximum off-diagonal value. 30

8.1 A taxonomy of the centrality measures discussed in this chapter, organised by the structural quantity each one fundamentally counts or aggregates. Measures actually computed on the real graph in Chapter 9 are marked with $\star$. . . . . 34

9.1 The eight highest-ranked zones by weighted degree centrality (full graph $\mathcal{G}$) and by eigenvector centrality (full graph $\mathcal{G}$), and the eight highest by betweenness centrality (sparsified $4$-NN graph $\mathcal{G}_4$). Values are not directly comparable across columns (different normalisations); rank order within each column is the substantive quantity. . . 35

11.1 Louvain community membership on the full thirty-three-zone graph $\mathcal{G}$ (random seed $42$; modularity $Q = 0.2290$). Four of the five communities correspond exactly to a single regional volume; the fifth merges three entire regional volumes into a single fifteen-zone community. . 44

12.1 The thirty-two declared edges of $\mathcal{G}_{\text{geo}}$, all intra-regional. No inter-regional land border exists among the thirty-three zone referents under the rule of Section 12.1. . . . . 48

12.2 Connected components of $\mathcal{G}_{\text{geo}}$ (seven total). Six coincide exactly with a regional volume; the British Isles volume splits into two geographic islands. . . . . 49

12.3 Agreement between three partitions of the thirty-three-zone vertex set: the six-region (regional-volume) partition, the seven-component geographic partition of Table 12.2, and the five-community full-graph phonotactic Louvain partition of Table 11.1. . . . . 50

13.1 The five triangles of $\mathcal{G}_{\text{geo}}$, each entirely within a single region. No triangle spans two regions, consistent with the absence of any declared inter-regional edge. . . . . 53

13.2 Diameter and average shortest-path length within each connected component of $\mathcal{G}_{\text{geo}}$ (component sizes and membership as in Table 12.2). Average path length is computed over all $\binom{n}{2}$ reachable pairs within the component; no cross-component pair is included, since none is connected. . . . . 54

15.1 The eight inter-regional bridge edges selected by the minimum spanning tree, in increasing order of weight. These are the minimal-cost connections the algorithm selects to link the six regional sub-trees into a single connected backbone. . . . . 59
15.2 The five region-level merge events of the formalized cross-volume dendrogram, derived by applying Kruskal's rule to Table 15.1 at region granularity. Heights are the real bridge weights of Table 15.1, unaltered. The remaining three bridges (Ireland–Georgian, Wales–Caucasian-TR, Alpine-Celtic–Azerbaijani) are redundant at this granularity, each reconnecting two regions already merged by an earlier, lower-height event. . . . . 61
15.3 Sensitivity of the full-graph Louvain partition to the proxy scaling parameter $\gamma$. ARI: Adjusted Rand Index against the six-region geographic ground truth. "AC–BI merge": whether at least one Anatolia–Caucasus zone and one British Isles zone are assigned to the same community. . . . . 63

16.1 Complexity of every algorithm used in this volume, with the practical cost at the present $|\mathcal{V}| = 33$, $|\mathcal{E}| = 528$ scale and the asymptotic bottleneck for a hypothetical much larger future application. . . . . 69

21.1 The thirty-three zones and their GeoNames source files, reproduced for reference (admin1 codes as documented in *Toponomastic Entropy*, Vols. II–VII). . . . . 86
21.2 Recomputed intra-regional symmetrised KL divergence (this chapter's independent pipeline) against the values inherited from the regional volumes (Table 7.2). All values in bits. . . . . 87

22.1 The twenty country-level vertices: P+T toponym pool size after deduplication, sample size, plug-in Shannon bigram entropy $H_2$, and Miller–Madow bias-corrected entropy $\hat{H}_2^{\text{MM}}$ (bits), computed directly from the supplied GeoNames files. . . . . 93
22.2 Eigenvector centrality, full ranking, country-level graph $G_{\text{country}}$. . . . 94
22.3 Betweenness centrality, top five non-zero vertices, country-level graph (true-distance weighting). . . . . 95
22.4 Minimum spanning tree of $G_{\text{country}}$ (true-distance weighting), all nineteen edges, lightest first. . . . . 97

24.1 Country-level pipeline recomputed under four values of the Lidstone smoothing parameter $\alpha$. $\bar{\delta}$, min, max: summary statistics of the 190-pair divergence matrix (bits). Top-3: the three highest eigenvector-centrality countries. $k$: number of Louvain communities. $Q$: modularity. 105

25.1 Five highest- and five lowest-Jaccard country pairs (bigram type-set overlap, $n = 3{,}000$ samples). . . . . 108

28.1 All 190 country pairs: great-circle distance between corpus centroids (km) and symmetrised KL divergence $\delta_{\text{sym}}$ (bits). . . . . 121

28.2 Top six positively distinctive bigrams per country (highest signed log-likelihood ratio $G^2$ against the pooled twenty-country base rate), in descending order of $G^2$. A leading ^ denotes a word-initial bigram and a trailing $ a word-final bigram (boundary markers as in Section 22.1); ␣ denotes a literal space character within a multi-word toponym. . . . 128

28.3 Plug-in Shannon bigram entropy $H_2$ (bits) as a function of sample size $n$, all twenty countries, identical pipeline to Section 22.1 other than the value of $n$. . . . . 130

28.4 Pearson correlation between the 190-entry country-level divergence matrix at sample size $n$ and the $n = 3{,}000$ reference matrix used in Chapter 22. . . . . 132

28.5 Observed Mantel correlation $r_{\text{obs}}$ between geographic distance and phonotactic divergence, as a function of sample size $n$. Compare Section 23.2, which reports the $n = 3{,}000$ value together with its permutation $p$-value. . . . . 133

28.6 Index of principal notation. . . . . 134

# List of Figures

7.1 Histogram of the $528$ edge-weight values $\delta(i, j)$ in the full thirty-three-zone graph (bin width $0.5$ bits), with intra-regional ($n = 75$, real KL-derived, mean $1.596$ bits) and inter-regional ($n = 453$, proxy, mean $2.756$ bits) edges shown in separate colours. The visible overlap region around $1.5$–$3.0$ bits, where both edge classes contribute substantially, confirms that edge weight alone cannot reliably distinguish edge type in this range; the heavy right tail of the intra-regional distribution ($\delta > 5$ bits) is entirely attributable to the Anatolia–Caucasus sub-graph. . 27

9.1 Eigenvector centrality, ten lowest-ranked zones and four highest-ranked zones (full graph $\mathcal{G}$), illustrating the steep drop from the Iberian Romance hub (Arabic, Castilian $> 0.51$) to the long tail of Anatolia–Caucasus and Central European peripheral zones ($< 0.10$). The gap between rank 10 (Illyrian-HR, $0.100$) and rank 4 (Galician, $0.197$) spans nearly the entire remaining range, indicating a markedly non-uniform centrality distribution rather than a smooth gradient. . 37

15.1 Minimum spanning tree of the full thirty-three-zone graph $\mathcal{G}$. Teal edges: intra-regional MST connections (real KL-derived, $24$ edges). Orange edges: the eight inter-regional bridge edges of Table 15.1, the minimal set of proxy-derived connections required for global connectivity. Georgian (Ge) is the only vertex with two inter-regional bridge edges, connecting both to Illyrian-SI (Central Europe) and to Ireland (British Isles). Two-letter zone abbreviations follow the order of Table 7.1 within each region. . . . . 60

15.2 The formalized cross-volume dendrogram: a six-leaf single-linkage tree over the regional volumes (II–VII), built by applying Kruskal's rule to the real bridge weights of Table 15.1 at region granularity. Merge heights are the unaltered real values from Table 15.2; the apparent tie at $1.554$ (steps 4 and 5) is drawn at two slightly offset heights purely for visual legibility and is noted explicitly as a true numerical tie in the underlying data. . . . . 61

21.1 Percolation curves for the $k$-nearest-neighbour ($k = 4$) sparsification of the recomputed $G'_{\text{phon}}$: fraction of nodes in the largest connected component as nodes are removed, under three removal orders. . . . . 90

22.1 Symmetrised KL divergence $\delta_{\text{sym}}(i, j)$ between every pair of the twenty country-level bigram distributions, computed directly from the supplied GeoNames files (darker = more divergent). The block structure visible along the diagonal — a Western-European block (ES–IE), a Maghreb block (MA, DZ, TN, MR), and a more diffuse South-Asian/Caucasus block — anticipates the Louvain partition of Section 22.4. . . . . 100
22.2 Force-directed (spring) layout of $G_{\text{country}}$, edges drawn from the union of the minimum spanning tree (Section 22.5) and each vertex's three nearest neighbours by $\delta_{\text{sym}}$, node colour by Louvain community membership (Table in Section 22.4). Layout computed by the Fruchterman–Reingold force-directed algorithm on affinity-weighted edges, random seed 42. . . . . 101
22.3 Percolation curves for the $k$-nearest-neighbour ($k = 4$) sparsification of $G_{\text{country}}$: fraction of nodes in the largest connected component as nodes are removed, under three removal orders, recomputed independently at country granularity following the methodology of Section 21.6. 101

23.1 Country-pair geographic distance (corpus centroids, thousands of km) plotted against symmetrised KL divergence (bits), all 190 pairs. The dashed line is an ordinary least-squares fit shown for visual reference only — the formal significance test reported in the text is the permutation-based Mantel test, not this OLS fit, since the 190 points are not independent observations. . . . . 103

28.1 Plug-in bigram entropy versus sample size $n$ for four representative countries, one per Louvain community of Section 22.4: Spain (Western-European cluster), the United Kingdom (Western-European cluster, insular branch), Algeria (Maghreb cluster), and India (South-Asian/Caucasus cluster). All four curves are monotonically increasing and visibly flattening by $n = 3{,}000$, the sample size used throughout Chapter 22. . . . . 132

# Preface

The present volume opens a second series, companion to *Toponomastic Entropy* (Vols. I–VII), which approaches the same empirical material — phonotactic distributions extracted from GeoNames toponymic corpora across thirty-three substrate zones in seven world regions — from a structurally different mathematical standpoint. Where the parent series treats each substrate zone as an isolated probability distribution and asks how *internally* diverse or *externally* divergent that distribution is (via Shannon entropy and Kullback–Leibler divergence taken pairwise), the present series asks a question that the pairwise framework cannot, by construction, answer: what is the *global relational structure* that emerges when all thirty-three zones are considered simultaneously as a single interconnected system?

This is a graph-theoretic question, and it requires a graph-theoretic apparatus: vertex degree and centrality measures, community detection algorithms, minimum spanning trees, and the spectral theory of graph Laplacians. None of this apparatus appears in the parent series, where the KL divergence matrices — though computed for every regional volume — are read row by row and column by column, never as the adjacency structure of a network.

**Relation to the parent series.** This volume is deliberately parasitic on *Toponomastic Entropy* in its empirical material: every KL divergence value used here to construct a graph edge is drawn, without modification, from the corresponding regional volume (Vol. II for Iberia, Vol. III for Central Europe, Vol. IV for the British Isles, Vol. V for the Indian Subcontinent, Vol. VI for Anatolia–Caucasus, Vol. VII for the Maghreb), cross-checked against the corrected master summary table of *Toponomastic Entropy*, Vol. I, Chapter 11. No new corpus sampling, bigram counting, or entropy estimation is performed in this volume; the present series is exclusively concerned with the question of what mathematical structure can be extracted from data that already exists.

**A methodological caveat stated at the outset.** The reader will notice that the six KL divergence matrices inherited from the parent series are each computed *within* a single regional corpus: every zone in the Iberian matrix was sampled, smoothed, and compared using the same protocol as every other Iberian zone, but no protocol exists for comparing an Iberian zone directly to a South Asian zone, because the parent series never performed that comparison. The thirty-three-zone graph constructed in Chapter 3 therefore necessarily distinguishes between two qualitatively different classes of edge: thirty-six *intra-regional* edges, each backed by a real, directly measured KL divergence value inherited from a regional volume, and four hundred ninety-two *inter-regional* edges, for which no directly measured divergence exists and which are instead assigned a declared *proxy distance* derived from the difference in bias-corrected Shannon entropy between the two zones. This proxy is

explicitly flagged wherever it is used, its limitations are discussed at length in Section 7.3, and every substantive structural claim in this volume is checked for robustness against the alternative of omitting inter-regional edges altogether (Chapter 15, Section 15.4).

This volume is self-contained for a reader with graduate-level training in probability and basic graph theory; familiarity with *Toponomastic Entropy*, Vol. I is recommended but not required, as all entropy and KL values used here are reproduced in context.

# Chapter 1

# From Pairwise Divergence to Network Structure

## 1.1 The Limits of the Pairwise Framework

The information-theoretic framework developed across the *Toponomastic Entropy* series treats the relationship between two substrate zones $i$ and $j$ as captured entirely by the (generally asymmetric) Kullback–Leibler divergence $D_{\mathrm{KL}}(\hat{p}_i \| \hat{p}_j)$ between their fitted bigram distributions. This is a sound and well-justified choice at the level of a *single pairwise comparison*: Chapter 4 of *Toponomastic Entropy*, Vol. I establishes in detail why KL divergence, rather than Euclidean or cosine distance, is the information-theoretically correct measure of dissimilarity between two categorical distributions over bigram space.

But a regional volume of the parent series typically reports five or six such zones and therefore $5 \times 4 = 20$ or $6 \times 5 = 30$ ordered pairwise divergence values, arranged in a matrix. The matrix is read — correctly, and with considerable interpretive insight in each regional volume — by scanning rows and columns for extreme values: the minimum off-diagonal cell, which identifies the most phonotactically proximate pair of zones (Etruscan–Venetian-Raetic in Central Europe, $D_{\mathrm{KL}} = 0.103$ bits; Ireland–Northern Ireland in the British Isles, $D_{\mathrm{KL}} = 0.205$ bits); the maximum, which identifies the most divergent pair (Caucasian-TR–Armenian in Anatolia–Caucasus, $D_{\mathrm{KL}} = 9.431$ bits); and the row sums, which identify zones whose corpus is, on average, poorly approximated by every other zone in the volume (the diagnostic underlying the Substrate Isolation Criterion of *Toponomastic Entropy*, Vol. I, Theorem 4.2).

This row-and-column reading strategy is a network-theoretic operation in disguise: a row sum is, up to normalisation, a weighted vertex degree: a "most isolated zone" claim is, in graph language, a low-centrality claim. The parent series performs these operations informally, region by region, without ever constructing the explicit graph object, computing a centrality measure with a name and a defined formula, or asking whether the apparent isolation of, say, Basque within the five-zone Iberian matrix is large or small *relative to the entire thirty-three-zone system*. A zone can be the most isolated of five without being particularly isolated in absolute terms; conversely, a zone of only moderate intra-regional isolation might, once embedded in the full cross-regional graph, turn out to occupy a structurally unique position. Answering this kind of question requires the graph to exist as a single object spanning

all thirty-three zones, not as six separate matrices.

## 1.2 What a Graph Adds

Treating the system of substrate zones as a weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$, with vertex set $\mathcal{V}$ the thirty-three zones and edge weights $w$ derived from KL divergence (Chapter 4 formalises the weighting scheme), opens three classes of question that the pairwise framework cannot pose:

1. **Centrality.** Which zones occupy structurally central positions in the network of phonotactic relationships, and which are peripheral? Three distinct notions of centrality — degree, betweenness, and eigenvector centrality — generally disagree, and the disagreement is itself diagnostic (Chapter 9).

2. **Community structure.** Does an algorithm that knows nothing about geography or language family, and sees only the weighted graph, recover the regional partition that was used to organise the parent series? Where it does not, what does the discrepancy reveal about cross-regional phonotactic affinities invisible to a region-by-region reading (Chapter 11)?

3. **Global connectivity and backbone structure.** What is the minimal set of edges that connects all thirty-three zones, and does that minimal backbone pass preferentially through certain "hub" zones — and does the spectral structure of the graph Laplacian (its eigenvalues) carry information about how fragmented or cohesive the overall system is (Chapter 15)?

**Remark: What This Volume Does Not Claim**

Network centrality and community detection are descriptive and exploratory tools. A zone with high betweenness centrality is not thereby shown to have played a historical role as a contact corridor; a Louvain community is not thereby shown to correspond to a real linguistic super-family. Every structural finding in this volume is reported as a property of the constructed graph, accompanied by an explicit discussion of which findings are robust to reasonable changes in graph construction choices (Section 15.4) and which are artefacts of those choices. The historical-linguistic *interpretation* of robust findings is offered as a hypothesis for further investigation, in the same prudent register as the parent series, not as an established result.

## 1.3 Outline of the Volume

Chapter 4 introduces the minimal graph-theoretic apparatus required: weighted graphs, adjacency and degree, paths and walks, the three centrality measures used throughout, and the spectral graph theory needed for Chapter 15. Chapter 7 constructs the thirty-three-node graph from the inherited KL matrices, states precisely the inter-regional proxy distance and its limitations, and reports basic descriptive statistics of the resulting graph. Chapter 9 computes and compares degree, betweenness, and eigenvector centrality for all thirty-three zones and interprets the disagreements. Chapter 11 applies Louvain community detection to the full graph and to

several robustness variants, and compares the discovered partition to the geographic regional partition inherited from the parent series. Chapter 15 examines global topological properties: the minimum spanning tree, the algebraic connectivity (Fiedler value) of the graph Laplacian, and a sensitivity analysis isolating which findings depend on the inter-regional proxy edges. Chapter 17 synthesises the findings, situates them relative to the parent series, and proposes an agenda for Volumes II–IV of the present series (differential geometry and topological data analysis on the space of toponymic forms; stochastic processes and Markov chains for areal diffusion; harmonic analysis and spectral methods for phonotactic sequence structure).

# Chapter 2

# Network Science in Linguistics: A Literature Review

Before building any apparatus of its own, this volume situates itself within an existing and substantial literature applying network-theoretic methods to language. The present volume is far from the first to model linguistic data as a graph; it differs from most prior work primarily in *which* linguistic object becomes a vertex (a geographic toponymic zone, rather than a word, a language, or a speaker) and in inheriting its edge weights directly from a prior, already-published information-theoretic analysis rather than fitting them from the graph construction stage itself. This chapter surveys the relevant strands of prior work and locates the present volume's specific contribution within them.

## 2.1 Word Co-occurrence and Semantic Networks

**Remark: The Earliest Strand: Words as Vertices**

The earliest and most extensively studied linguistic networks take individual words as vertices, with edges defined by co-occurrence, syntactic dependency, or semantic association. Ferrer i Cancho and Solé (2001) showed that word co-occurrence networks exhibit small-world and (approximately) scale-free degree distributions, properties this volume's Chapter 5 defines formally; Steyvers and Tenenbaum (2005) extended the approach to free-association and WordNet-derived semantic networks, finding the same broad small-world signature across three distinct semantic resources. This strand established, more than two decades ago, that graph-theoretic tools developed for technological and biological networks transfer productively to lexical structure – the foundational precedent this volume's much narrower toponymic application ultimately rests on.

## 2.2 Phonological and Phonotactic Networks

**Remark: The Strand Closest to This Volume's Subject Matter**

Closer to the present volume's specific concern with phonotactic structure, Vitevitch (2008) modelled phonological neighbourhood relations among English words (edges connecting word pairs differing by a single phoneme) as a network, finding that a word's network degree (its phonological neighbourhood density) predicts measurable behavioural outcomes in speech production and lexical access – a finding of direct relevance to this volume's own assumption (Chapter 4) that graph-structural position can carry psycholinguistically or historically meaningful information beyond what a word- or zone-level summary statistic alone reports. Arbesman et al. (2010) extended this to a comparative cross-linguistic study of phonological networks in English and Spanish, finding broadly similar small-world structure across the two languages despite substantial differences in phoneme inventory size and phonotactic constraints – suggestive evidence that network-level regularities can be more stable across languages than the surface phonological systems generating them.

## 2.3 Dialectometry and Areal Linguistics

**Remark: The Strand Closest to This Volume's Geographic Framing**

Dialectometry – the quantitative study of dialect variation as a function of geography – has used network and graph methods since at least Nerbonne (2010)'s review of the field's computational turn, though the dominant tools in that tradition (multidimensional scaling, hierarchical clustering, the Levenshtein-distance-based aggregation methods surveyed by Heeringa (2004)) are more closely aligned with the *Toponomastic Entropy* parent series' clustering apparatus than with this volume's explicit graph-theoretic reading. Where this volume differs is in formalising the geographic-versus-linguistic-distance relationship as a graph comparison problem in its own right (the geographic-adjacency-versus-phonotactic-affinity contrast that motivates Chapter 7's inter-regional proxy construction), rather than treating geography purely as a covariate or a constraint on permissible clustering, the more common dialectometric framing. Wieling et al. (2011) applied network-style visualisation (specifically, edge-bundled graphs of pairwise dialect distance) to Dutch dialect data without adopting the formal graph-theoretic toolkit (centrality, modularity, spectral methods) this volume develops, illustrating a broader pattern in the dialectometric literature: graph *visualisation* of linguistic distance is well-established, while graph *analysis* in the sense of Chapter 4 remains comparatively rare.

## 2.4 Language Contact and Areal Diffusion as Network Phenomena

**Remark: The Strand Anticipating Volume III**

A smaller but directly relevant strand treats language contact and areal diffusion explicitly as network or graph-diffusion phenomena. Abrams and Strogatz (2003) modelled language competition and extinction dynamics using a system closely related to the random-walk and diffusion apparatus this series' Volume III develops in detail, applying it to bilingual competition between two languages in a shared population rather than to areal diffusion of phonotactic traits across zones, but sharing the underlying mathematical machinery (coupled differential equations admitting a random-walk-type stationary analysis). Croft (2000) argued, on purely qualitative grounds, for treating language change as a population-level diffusion process with strong structural analogies to epidemiological and ecological spread models – a position this volume's Chapter 17 explicitly does *not* adopt without qualification (the present series consistently distinguishes a graph's synchronic structure from any diachronic diffusion claim built on top of it, a distinction Volume III's preface develops at length), but one whose basic intuition (that contact-induced change should look, structurally, like diffusion on a network of populations) is exactly the intuition motivating Volume III's research programme.

## 2.5 What Is Genuinely New in This Volume's Approach

**Methodological Transparency: An Honest Positioning Statement**

None of the individual techniques surveyed in this chapter is new: betweenness and eigenvector centrality, modularity-based community detection, and spectral graph methods are all decades-old tools, already applied to lexical, phonological, and (less formally) dialectometric data by the literature just reviewed. What this volume contributes is narrower and more specific: the application of this established graph-theoretic toolkit, in a methodologically disciplined way (real intra-regional edge weights inherited unmodified from a prior published analysis; a declared, explicitly-flagged proxy for inter-regional edges; every claim about historical or diachronic meaning explicitly withheld pending Volume III), to the specific domain of toponymic substrate zones defined by *Toponomastic Entropy*'s thirty-three-zone partition. The contribution is one of careful, transparently-documented application and synthesis across two previously separate strands (the information-theoretic toponymics of the parent series and the network-theoretic toolkit surveyed here) rather than of new graph-theoretic technique.

# Chapter 3

# The Metric Structure of the Edge-Weight Function

Every shortest-path computation, every betweenness centrality calculation, and every minimum-spanning-tree construction in this volume silently assumes that $\delta(i, j)$, the edge-weight function inherited from the parent series, behaves enough like an ordinary geometric distance for these algorithms' usual properties (e.g. that a shortest path is genuinely shortest, that the triangle inequality bounds any detour) to hold. This assumption deserves to be checked rather than taken for granted, particularly because the parent series' underlying quantity – Kullback–Leibler divergence – is well known *not* to satisfy it in general. This chapter examines exactly what is, and is not, guaranteed.

## 3.1 Metric Axioms

**Definition: Metric**

A function $\delta : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_{\geq 0}$ is a metric if it satisfies, for all $i, j, k \in \mathcal{V}$: (M1) *non-negativity*, $\delta(i, j) \geq 0$; (M2) *identity of indiscernibles*, $\delta(i, j) = 0 \iff i = j$; (M3) *symmetry*, $\delta(i, j) = \delta(j, i)$; and (M4) the *triangle inequality*, $\delta(i, k) \leq \delta(i, j) + \delta(j, k)$. A function satisfying only (M1), (M3), (M4) but not requiring (M2) to hold as a strict biconditional (allowing $\delta(i, j) = 0$ for $i \neq j$) is a *pseudometric*; one satisfying (M1), (M2), (M4) but not (M3) is a *quasimetric*.

**Proposition: KL Divergence Is Neither Symmetric Nor Triangle-Inequality-Respecting**

For probability distributions $P, Q$, Kullback–Leibler divergence $D_{\mathrm{KL}}(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$ satisfies (M1) (non-negativity, by Gibbs' inequality) but generally fails both (M3) (it is well known and easily checked on simple two-outcome examples that $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$ in general) and (M4) (counterexamples exist where $D_{\mathrm{KL}}(P\|R) > D_{\mathrm{KL}}(P\|Q) + D_{\mathrm{KL}}(Q\|R)$ for suitably chosen intermediate $Q$, a consequence of $D_{\mathrm{KL}}$'s lack of any general sub-additivity property across distinct argument pairs).

**Remark: Why This Matters Directly for the Present Graph**

The parent series (*Toponomastic Entropy*) computes pairwise KL divergences between zones' bigram distributions, and this volume's graph edges inherit those values. If $\delta(i, j)$ in this volume were literally raw, directed $D_{\mathrm{KL}}(P_i \| P_j)$, none of the algorithms used in Chapters 9–15 (Dijkstra, Brandes' betweenness, Kruskal's MST) would be guaranteed to behave correctly, since all three implicitly assume (M3) symmetry at minimum (an undirected graph requires a single weight per edge, not two directed weights) and Kruskal's MST optimality proof additionally leans on the triangle inequality holding for the notion of "distance" being minimised. This is precisely why the parent series symmetrises its divergence (Chapter 3 of *Toponomastic Entropy*, Vol. I), and this volume inherits the symmetrised quantity, not raw directed KL.

## 3.2 Jensen–Shannon Divergence Is a True Metric

**Definition: Jensen–Shannon Divergence**

The Jensen–Shannon divergence symmetrises KL divergence via a shared midpoint distribution $M = \frac{1}{2}(P + Q)$: $\mathrm{JSD}(P, Q) = \frac{1}{2} D_{\mathrm{KL}}(P \| M) + \frac{1}{2} D_{\mathrm{KL}}(Q \| M)$. By construction $\mathrm{JSD}(P, Q) = \mathrm{JSD}(Q, P)$ (M3 holds), and $\mathrm{JSD}(P, Q) = 0 \iff P = Q$ (M2 holds, since $D_{\mathrm{KL}}(P \| M) = 0$ requires $P = M$, which combined with $M = \frac{1}{2}(P + Q)$ forces $P = Q$). JSD is also bounded, $0 \leq \mathrm{JSD}(P, Q) \leq \log 2$ (in nats), unlike unbounded raw $D_{\mathrm{KL}}$.

**Theorem: $\sqrt{\mathrm{JSD}}$ Is a Metric**

(Endres and Schindelin, 2003; Fuglede and Topsoe, 2004) The square root of Jensen–Shannon divergence, $\sqrt{\mathrm{JSD}(P, Q)}$, satisfies the full triangle inequality (M4) and is therefore a genuine metric on the space of probability distributions, even though JSD itself (without the square root) is not guaranteed to satisfy (M4) for every triple of distributions.

**Remark: A Methodological Note for This Series**

This theorem – that $\sqrt{\mathrm{JSD}}$, not raw JSD, is the genuine metric – is the formal justification for treating distances built from Jensen–Shannon divergence as triangle-inequality-respecting only after the square-root transform; readers checking the parent series' exact edge-weight convention should verify which of the two (JSD or $\sqrt{\mathrm{JSD}}$) was actually used before assuming any triangle-inequality-dependent algorithmic guarantee (Kruskal's MST optimality, in particular) applies without qualification. The present volume flags this explicitly rather than silently assuming the favourable case.

## 3.3 Why Triangle-Inequality Violations Would (and Would Not) Matter

**Proposition: Dijkstra Requires Only Non-Negativity, Not the Triangle Inequality**

Dijkstra's shortest-path algorithm's correctness proof relies on (M1) non-negativity alone (the greedy frontier-expansion argument requires only that adding an edge never *decreases* a path's total weight, which non-negativity guarantees); it does not require (M4) the triangle inequality to hold for the underlying weight function, since Dijkstra computes shortest *paths* directly rather than relying on any prior assumption about what the direct edge weight between two non-adjacent vertices "should" be.

**Methodological Transparency: Kruskal's MST Optimality Does Lean on a Distance Interpretation**

Kruskal's algorithm is provably correct (it computes a true minimum spanning tree) for *any* real-valued edge-weight function, with or without the triangle inequality – the greedy exchange argument underlying its optimality proof (Cormen et al., 2009) requires only that edge weights are real numbers admitting a total order, not that they satisfy any metric axiom. What *does* depend on the triangle inequality is the *interpretation* of the resulting tree as a meaningful approximation to some notion of geometric closeness: without (M4), a path through several MST edges could, in principle, connect two zones far more cheaply than their (hypothetical) direct edge would suggest, undermining any claim that MST path length is a sensible proxy for "how close" two zones are along the tree. This is precisely why Section 2's verification that the present volume's edge weights derive from $\sqrt{\text{JSD}}$-type quantities (rather than raw asymmetric $D_{\text{KL}}$) matters beyond pure formal tidiness: it is what licenses reading MST path length, as Volume III does throughout, as a meaningful distance rather than merely an artefact of whichever greedy tree the algorithm happened to construct.

# Chapter 4

# Graph-Theoretic Foundations

This chapter assembles the minimal apparatus of graph theory required for the remainder of the volume. Readers already familiar with weighted graphs, centrality measures, and spectral graph theory may proceed directly to Chapter 7; the treatment here is included for self-containedness and to fix notation.

## 4.1 Weighted Graphs

**Definition: Weighted Graph**

A weighted, undirected graph is a triple $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$, where $\mathcal{V}$ is a finite set of vertices, $\mathcal{E} \subseteq \binom{\mathcal{V}}{2}$ is a set of unordered vertex pairs (edges), and $w : \mathcal{E} \to \mathbb{R}_{>0}$ assigns a positive weight to each edge. The graph is *complete* if $\mathcal{E} = \binom{\mathcal{V}}{2}$, i.e. every pair of vertices is connected.

In the present application, $\mathcal{V}$ is the set of thirty-three substrate zones, and the edge weight $w(i, j)$ is derived from a measure of phonotactic similarity (Section 7.2) between zones $i$ and $j$. Two conventions for $w$ recur throughout the volume and must be kept distinct:

- A **distance-like** weight, denoted $\delta(i, j)$, for which *larger* values indicate *less* similar zones. The symmetrised KL divergence $\delta(i, j) = \tfrac{1}{2}\big[D_{\mathrm{KL}}(\hat{p}_i \| \hat{p}_j) + D_{\mathrm{KL}}(\hat{p}_j \| \hat{p}_i)\big]$, already used for hierarchical clustering in *Toponomastic Entropy*, Vol. I, Section 5.3, is the natural distance-like weight here.

- An **affinity-like** weight, denoted $a(i, j)$, for which *larger* values indicate *more* similar zones. Affinity weights are required by community-detection algorithms (Louvain modularity maximisation is defined for affinity, not distance, graphs) and by eigenvector centrality. We use $a(i, j) = 1/(\delta(i, j) + \varepsilon)$ with a small regularisation constant $\varepsilon = 0.05$ bits to avoid division by zero on the (here non-occurring) case of exactly zero divergence.

Distance-like and affinity-like weights are used for different operations throughout this volume: shortest paths, betweenness centrality, and minimum spanning trees use $\delta$; community detection and eigenvector centrality use $a$. The choice is not arbitrary cosmetics but reflects which graph operation is well-defined for which

weight type: Dijkstra's shortest-path algorithm presupposes that summing edge weights along a path produces a meaningful notion of total "cost", which holds for $\delta$ (KL divergences add along a chain of approximations, in the spirit of the data-processing inequality) but not for $a$.

## 4.2 Degree and Local Structure

**Definition: Weighted Degree**

The weighted degree of vertex $i$ in a graph with affinity weights $a$ is

$$\deg(i) \;=\; \sum_{j \neq i} a(i,j).$$

The *degree centrality* of vertex $i$ is the normalised quantity $C_{\mathrm{D}}(i) = \deg(i)/(|\mathcal{V}|-1)$, dividing by the maximum possible degree in a graph on $|\mathcal{V}|$ vertices.

**Remark: A Trivial Case to Avoid**

On a *complete* graph (every vertex connected to every other), degree centrality is uninformative as a discriminator between vertices unless edge weights vary substantially, since every vertex has the same number of edges $|\mathcal{V}| - 1$ by construction. The thirty-three-zone graph constructed in Chapter 7 is complete once inter-regional proxy edges are included (Section 7.3), so weighted degree, not the unweighted vertex degree, is the relevant quantity throughout; an early diagnostic check in Section 7.4 confirms that weighted degree varies enough across the thirty-three zones to be informative.

## 4.3 Paths, Geodesics, and Betweenness Centrality

**Definition: Shortest Path and Geodesic Distance**

For a distance-weighted graph $(\mathcal{V}, \mathcal{E}, \delta)$, a path from $i$ to $j$ is a sequence of vertices $i = v_0, v_1, \ldots, v_k = j$ with $(v_{l-1}, v_l) \in \mathcal{E}$ for all $l$. Its length is $\sum_{l=1}^{k} \delta(v_{l-1}, v_l)$. The geodesic distance $d(i,j)$ is the minimum length over all paths from $i$ to $j$.

On a complete graph the geodesic distance trivially coincides with the direct edge weight $\delta(i,j)$, since the direct edge is itself a path of length one and cannot be improved upon by a multi-hop detour through positive weights. This observation is not vacuous: it means that on the complete thirty-three-zone graph, *shortest-path* based diagnostics (geodesic distance, and closeness centrality, which is defined from it) reduce to the direct pairwise distances already present in the inherited KL matrices, and add no new information beyond re-deriving what is already in *Toponomastic Entropy*. The graph-theoretic apparatus earns its keep instead on *sparsified* or *tree-structured* derived graphs — the minimum spanning tree of Chapter 15 foremost among them — where genuine multi-hop paths exist and geodesic distance can differ substantially from direct distance.

**Definition: Betweenness Centrality**

The betweenness centrality of vertex $v$ is

$$C_{\mathrm{B}}(v) \;=\; \sum_{\substack{i,j\in\mathcal{V}\\ i\neq j\neq v}} \frac{\sigma_{ij}(v)}{\sigma_{ij}},$$

where $\sigma_{ij}$ is the number of shortest (geodesic) paths between $i$ and $j$, and $\sigma_{ij}(v)$ is the number of those paths passing through $v$. Intuitively, $C_{\mathrm{B}}(v)$ measures how often $v$ lies on the shortest route between two other vertices.

On the complete graph, every shortest path is the direct edge (as just noted), so $\sigma_{ij}(v) = 0$ for every $v \neq i, j$ and consequently $C_{\mathrm{B}} \equiv 0$ for all vertices: betweenness centrality is *identically zero and uninformative* on the raw complete graph. This is not a defect of the framework but a structural fact that dictates methodology: meaningful betweenness centrality, in the present application, can only be computed on a *sparsified* version of the graph (e.g. a $k$-nearest-neighbour graph, or the minimum spanning tree), where genuine multi-hop shortest paths exist. Chapter 9 computes betweenness on a sparsified graph constructed for this purpose and states the sparsification rule explicitly.

## 4.4 Eigenvector Centrality

**Definition: Eigenvector Centrality**

Let $A$ be the weighted adjacency matrix of $\mathcal{G}$ with affinity weights, $A_{ij} = a(i, j)$ for $(i, j) \in \mathcal{E}$ and $A_{ij} = 0$ otherwise. The eigenvector centrality vector $\mathbf{x}$ is the (entrywise non-negative, by the Perron–Frobenius theorem) eigenvector associated with the largest eigenvalue $\lambda_{\max}$ of $A$:

$$A\mathbf{x} = \lambda_{\max}\,\mathbf{x}, \qquad x_i \geq 0\ \forall i.$$

Vertex $i$'s eigenvector centrality $x_i$ is large precisely when $i$ is strongly connected (large $a(i, j)$) to other vertices that are themselves highly central — a recursive, self-consistent notion of importance distinct from the purely local degree centrality.

**Remark: Why Eigenvector Centrality Is the Right Tool Here**

Degree centrality on the complete weighted graph measures total affinity summed over all thirty-two neighbours; eigenvector centrality re-weights that sum by the importance of the neighbour itself. A zone strongly linked (high KL-derived affinity) only to otherwise peripheral zones receives a smaller eigenvector centrality than one linked, even slightly less strongly in raw affinity, to zones that are themselves hubs. Because thirty-three zones with a single, declared affinity weighting is exactly the regime in which this recursive distinction is most likely to produce non-trivial, non-degree-redundant information, eigen-

vector centrality — unlike betweenness — remains well-defined and informative on the raw complete graph and is used as such in Chapter 9.

## 4.5 Community Detection: Modularity and the Louvain Algorithm

**Definition: Modularity**

Given a partition $\mathcal{C} = \{C_1, \ldots, C_k\}$ of $\mathcal{V}$ into disjoint communities, the modularity of the partition with respect to affinity weights $a$ is

$$Q(\mathcal{C}) \;=\; \frac{1}{2m} \sum_{i,j} \left[ a(i,j) - \frac{\deg(i)\deg(j)}{2m} \right] \mathbb{1}[c_i = c_j],$$

where $m = \frac{1}{2}\sum_{i,j} a(i,j)$ is the total edge weight, $c_i$ is the community to which $i$ is assigned, and $\mathbb{1}[\cdot]$ is the indicator function. Modularity compares the observed within-community edge weight to the weight expected under a null model that preserves each vertex's degree but rewires edges uniformly at random; positive modularity indicates more within-community weight than this null model predicts.

The Louvain algorithm (Blondel et al., 2008) greedily maximises $Q(\mathcal{C})$ via a two-phase iterative procedure (local vertex-moving to maximise modularity gain, followed by community aggregation into super-vertices, repeated until convergence). It is the de facto standard community-detection algorithm for weighted graphs of this size and is used throughout Chapter 11, together with a stated random seed for reproducibility (modularity optimisation is NP-hard in general, and Louvain's greedy local-search nature means results can depend weakly on traversal order).

**Methodological Transparency: Modularity Is Sensitive to Graph Density**

A known property of modularity-based community detection, sometimes called the *resolution limit* (Fortunato, 2010), is that modularity values are not directly comparable across graphs of different density or size, and that very dense or very sparse graphs can produce spuriously high or spuriously fragmented community structure regardless of any genuine underlying signal. Because the raw thirty-three-zone graph constructed in Chapter 7 is complete (every pair of zones connected, whether by a real or a proxy edge), and because intra-regional edges carry systematically smaller $\delta$ (hence larger $a$) than inter-regional proxy edges by construction (Section 7.3), a naive application of Louvain to this graph risks simply rediscovering the six geographic regions as an artefact of the graph-construction procedure rather than as a genuine cross-regional finding. Chapter 11 addresses this directly by reporting modularity and community membership both on the full graph and on the intra-regional-only graph, and by interpreting agreement or disagreement between the two as the substantive finding.

## 4.6 Spectral Graph Theory: The Graph Laplacian

**Definition: Graph Laplacian and Algebraic Connectivity**

For a weighted graph with affinity adjacency matrix $A$ and degree matrix $D = \text{diag}(\deg(1), \ldots, \deg(|\mathcal{V}|))$, the (combinatorial) graph Laplacian is $L = D - A$. $L$ is symmetric and positive semi-definite; its eigenvalues $0 = \mu_1 \leq \mu_2 \leq \cdots \leq \mu_{|\mathcal{V}|}$ are real and non-negative. The multiplicity of the eigenvalue $0$ equals the number of connected components of the graph. The second-smallest eigenvalue $\mu_2$, known as the *algebraic connectivity* or *Fiedler value* (Fiedler, 1973), quantifies how well-connected the graph is as a whole: larger $\mu_2$ indicates a more cohesive, harder-to-disconnect graph; $\mu_2$ near zero indicates a graph close to being disconnected into separate components.

The Fiedler value and its associated eigenvector (the *Fiedler vector*) are used in Chapter 15 as a global, single-number summary of how fragmented the thirty-three-zone system is, and as a spectral bisection tool that partitions vertices by the sign of their Fiedler-vector entry — a partition method conceptually distinct from, and useful as a cross-check against, Louvain modularity maximisation.

## 4.7 Minimum Spanning Trees

**Definition: Minimum Spanning Tree**

For a connected, distance-weighted graph $(\mathcal{V}, \mathcal{E}, \delta)$, a spanning tree is a connected, acyclic subgraph $T = (\mathcal{V}, \mathcal{E}_T)$ with $\mathcal{E}_T \subseteq \mathcal{E}$ and $|\mathcal{E}_T| = |\mathcal{V}| - 1$. The minimum spanning tree (MST) is the spanning tree minimising $\sum_{(i,j)\in\mathcal{E}_T} \delta(i,j)$. Kruskal's and Prim's algorithms both compute an MST in polynomial time; the present volume uses Kruskal's algorithm as implemented in standard graph libraries.

The MST plays a distinguished methodological role in this volume: it is the unique minimal "backbone" connecting all thirty-three zones using the fewest and lightest possible edges, and consequently it is the sparsified graph on which betweenness centrality (Section above) becomes non-trivial. Moreover, examining which edges of the MST cross a regional boundary — i.e. which inter-regional proxy edges turn out to be *indispensable* for global connectivity at minimum total distance — provides one of the clearest diagnostics in the volume for which cross-regional phonotactic affinities are structurally load-bearing rather than incidental (Chapter 15, Section 15.1).

# Chapter 5

# Extended Graph-Theoretic Toolkit

The previous chapter assembled the minimal apparatus needed to state what this volume's central questions even mean: degree, paths, betweenness, eigenvector centrality, modularity, the Laplacian, the minimum spanning tree. That apparatus is sufficient to carry out the analyses of Chapters 7–15, but it leaves several natural follow-up questions unanswered – questions a careful reader of network science would immediately ask, and that this chapter addresses before the thirty-three-zone graph is actually built. None of the material here introduces new empirical claims about the toponymic data; every numerical illustration in this chapter uses small, explicitly synthetic toy graphs (five or six vertices, declared as such) chosen purely to make a definition concrete, exactly the discipline Volume IV later applies to its synthetic sequences.

**Remark: Why This Chapter Exists**

A reader who knows only the five concepts of Chapter 4 might reasonably ask: how clustered is the network locally? How far apart are two typical zones? Does high-degree tend to connect to high-degree (assortative) or to low-degree (disassortative)? How much of what the graph shows is simply what *any* graph with this many edges would show, regardless of which specific pairs are connected? This chapter supplies the tools to ask these questions of the real thirty-three-zone graph once it is built in Chapter 7, and several of its constructions (the configuration-model null, in particular) are used directly in Chapter 15's robustness analysis.

## 5.1 Local Clustering: The Clustering Coefficient

**Definition: Local Clustering Coefficient**

For a vertex $i$ with neighbour set $N(i) = \{j : (i,j) \in \mathcal{E}\}$ and degree $k_i = |N(i)|$, the local clustering coefficient is

$$C_i \;=\; \frac{2\,|\{(j,k) : j,k \in N(i),\ (j,k) \in \mathcal{E}\}|}{k_i(k_i-1)}, \qquad k_i \geq 2,$$

the fraction of $i$'s neighbour pairs that are themselves connected – i.e. the fraction of possible "triangles through $i$" that are actually closed. For $k_i < 2$,

$C_i$ is conventionally undefined (or set to $0$ by convention, the choice adopted here). The graph-level *average* clustering coefficient is $\bar{C} = \frac{1}{|\mathcal{V}|}\sum_i C_i$ (Watts–Strogatz's original definition), to be distinguished from the *global* (transitivity) coefficient $C_{\text{glob}} = 3 \times (\text{number of triangles})/(\text{number of connected triples})$, which weights high-degree vertices more heavily since they participate in more connected triples.

**Proposition: $\bar{C}$ and $C_{\text{glob}}$ Generally Disagree**

$\bar{C}$ and $C_{\text{glob}}$ coincide only in the special case where every vertex has the same degree (a regular graph); in general $C_{\text{glob}}$ is a degree-weighted average of the $C_i$'s, with weight proportional to $k_i(k_i - 1)$, so a graph with a few very-high- degree hubs whose neighbourhoods are sparse will show $C_{\text{glob}} \ll \bar{C}$, while $\bar{C}$ is dominated by the (possibly many) low-degree vertices, for whom a single triangle already gives $C_i = 1$.

**Example: Clustering on a Five-Vertex Toy Graph**

Take the illustrative (non-empirical) graph on vertices $\{a, b, c, d, e\}$ with edges $\{ab, ac, bc, cd, de\}$ – a triangle $abc$ with a pendant path $c$–$d$–$e$ attached. Then $k_a = k_b = 2$, $k_c = 3$, $k_d = 2$, $k_e = 1$. Vertex $a$'s neighbours are $\{b, c\}$, and $bc \in \mathcal{E}$, so $C_a = 2 \cdot 1/(2 \cdot 1) = 1$; by symmetry $C_b = 1$. Vertex $c$'s neighbours are $\{a, b, d\}$; among these pairs only $ab \in \mathcal{E}$, so $C_c = 2 \cdot 1/(3 \cdot 2) = 1/3$. Vertex $d$'s neighbours are $\{c, e\}$, and $ce \notin \mathcal{E}$, so $C_d = 0$; $C_e$ is undefined ($k_e = 1$). Averaging over the four defined vertices, $\bar{C} = (1 + 1 + \frac{1}{3} + 0)/4 = 7/12 \approx 0.583$. The graph has exactly one triangle ($abc$) and $\binom{2}{1} + \binom{3}{2} + \binom{2}{1} + \binom{1}{0} = 2+3+2+0 = 7$ connected triples through some vertex (counting each triple once at its centre), so $C_{\text{glob}} = 3 \cdot 1/7 \approx 0.429$ – illustrating Proposition above: the two coefficients disagree because degree is heterogeneous.

## 5.2 Distance Structure: Diameter, Radius, and Average Path Length

**Definition: Diameter, Radius, and Average Path Length**

For a connected weighted graph with shortest-path distance $d(i, j)$ between vertices $i$ and $j$ (the sum of edge weights along the lightest path, computable by Dijkstra's algorithm), the *eccentricity* of $i$ is $\text{ecc}(i) = \max_j d(i, j)$, the *diameter* is $\text{diam}(\mathcal{G}) = \max_i \text{ecc}(i)$, the *radius* is $\text{rad}(\mathcal{G}) = \min_i \text{ecc}(i)$, and the *average path length* is $\bar{d} = \binom{|\mathcal{V}|}{2}^{-1} \sum_{i<j} d(i, j)$. A vertex attaining the radius is called a *graph centre*; it need not be unique.

**Proposition: The Radius–Diameter Inequality**

For any connected graph, $\text{rad}(\mathcal{G}) \le \text{diam}(\mathcal{G}) \le 2\,\text{rad}(\mathcal{G})$.

*Proof.* The first inequality holds because the centre's own eccentricity is, by definition, at least the minimum eccentricity, i.e. $\mathrm{rad}(\mathcal{G})$ is itself an eccentricity value and the diameter is the maximum eccentricity, so rad $\leq$ diam trivially when $|\mathcal{V}| > 1$. For the second: let $c$ be a centre, so $\mathrm{ecc}(c) = \mathrm{rad}$, and let $u, v$ be any pair with $d(u, v) = \mathrm{diam}$. By the triangle inequality for shortest-path distance (itself a consequence of path concatenation: any walk from $u$ to $c$ to $v$ is a valid, possibly non-shortest, $u$–$v$ walk, so $d(u, v) \leq d(u, c) + d(c, v)$), we have $\mathrm{diam} = d(u, v) \leq d(u, c) + d(c, v) \leq \mathrm{ecc}(c) + \mathrm{ecc}(c) = 2\,\mathrm{rad}$. □

**Example: Diameter on the Same Toy Graph**

On the illustrative path-plus-triangle graph of the previous example, with all edge weights set to 1 for simplicity: $d(a, e) = d(a, c) + d(c, d) + d(d, e) = 1 + 1 + 1 = 3$ is the largest pairwise distance, so $\mathrm{diam} = 3$; vertex $c$ has $\mathrm{ecc}(c) = \max(1, 1, 1, 2) = 2$, the smallest of any vertex's eccentricity (checking $a$: $\mathrm{ecc}(a) = \max(1, 1, 2, 3) = 3$; $d$: $\mathrm{ecc}(d) = \max(2, 2, 1, 1) = 2$; so $c$ and $d$ tie for centre), so $\mathrm{rad} = 2$, and indeed $2 \leq 3 \leq 4$ as the Proposition requires.

## 5.3 Degree Assortativity

**Definition: Assortativity Coefficient (Newman 2002)**

The degree assortativity coefficient is the Pearson correlation coefficient of the degrees at the two ends of each edge, computed over all edges (each edge contributing both orderings):

$$r = \frac{\sum_{(i,j)\in\mathcal{E}} k_i k_j / |\mathcal{E}| - \left[\sum_{(i,j)\in\mathcal{E}} \tfrac{1}{2}(k_i + k_j)/|\mathcal{E}|\right]^2}{\sum_{(i,j)\in\mathcal{E}} \tfrac{1}{2}(k_i^2 + k_j^2)/|\mathcal{E}| - \left[\sum_{(i,j)\in\mathcal{E}} \tfrac{1}{2}(k_i + k_j)/|\mathcal{E}|\right]^2} \in [-1, 1].$$

A graph is *assortative* ($r > 0$) if high-degree vertices tend to connect to other high-degree vertices, and *disassortative* ($r < 0$) if high-degree vertices tend to connect to low-degree ones. Newman (2002) observed that most social networks are assortative while most technological and biological networks are disassortative – a contrast whose explanation (homophily versus engineering constraints on hub fan-out) is itself a long-running topic in the network science literature, and one this volume revisits once the real thirty-three-zone assortativity is computed in Chapter 7.

**Remark: Why a Complete Weighted Graph Complicates Assortativity**

The classical assortativity coefficient is defined for simple (unweighted, single-edge) graphs; the present volume's graph is complete (every pair of zones is connected) and weighted, so an unweighted-degree notion of assortativity is not directly meaningful (every vertex has the same unweighted degree, $|\mathcal{V}| - 1 = 32$, making $r$ trivially undefined by zero variance in the denominator). The relevant adaptation, used in Chapter 7, replaces degree with *weighted* degree (strength)

$s_i = \sum_j w(i,j)$ and computes the analogous correlation on edge *weights* versus the strengths of their endpoints – a standard generalisation (Barrat et al., 2004) for exactly this situation.

## 5.4 Random Graph Null Models

A recurring methodological question in this volume (sharpened further in Chapters 7 and 15) is: how much of an observed structural feature is special to the real toponymic data, and how much would *any* graph with comparable basic statistics (number of vertices, number of edges, degree sequence) show? Answering this requires a family of null models against which to compare.

**Definition: Erdős–Rényi Random Graph $G(n,p)$**

The Erdős–Rényi model $G(n,p)$ places an edge between each of the $\binom{n}{2}$ vertex pairs independently with probability $p$. The expected number of edges is $p\binom{n}{2}$, the expected degree of every vertex is $p(n-1)$ (so the degree distribution concentrates around its mean for large $n$, by a standard concentration argument: each vertex's degree is a sum of $n-1$ i.i.d. Bernoulli$(p)$ variables, to which Chernoff bounds apply), and the expected clustering coefficient is exactly $p$ (since, conditional on $i$ having two neighbours $j,k$, the edge $jk$ is present with the same unconditional probability $p$ by independence).

**Definition: Configuration Model**

Given a target degree sequence $(k_1,\ldots,k_n)$, the configuration model generates a random graph (technically, a random multigraph, though self-loops and multi-edges become asymptotically rare for "nice" degree sequences) with exactly that degree sequence by giving each vertex $i$ a set of $k_i$ *half-edges* ("stubs"), then choosing a uniformly random perfect matching of all stubs into edges. This is the standard null model for asking "is feature $X$ of the real graph explained simply by its degree sequence, or does it require the specific identity of which vertices connect to which?" – precisely the logic Chapter 7's Section on basic descriptive statistics and Chapter 15's permutation tests both instantiate in a weighted-graph adaptation.

**Proposition: $G(n,p)$ Has Vanishing Clustering at Fixed Mean Degree**

If $p = c/n$ for fixed constant $c$ (so the expected degree $c$ stays fixed as $n \to \infty$, the regime relevant to sparse real-world networks), then the expected clustering coefficient $p = c/n \to 0$.

*Proof.* Immediate from the previous defbox: $\mathbb{E}[C] = p$ exactly under $G(n,p)$, and $p = c/n \to 0$ as $n \to \infty$ for fixed $c$. □

**Remark: Why This Matters for a 33-Vertex Complete Graph**

The present volume's graph is complete ($p = 1$ in $G(n, p)$ terms), so the Erdős–Rényi clustering result is not directly diagnostic here (a complete graph trivially has $C_i = 1$ for every vertex). It becomes directly relevant, however, the moment a *sparsified* graph is considered – the minimum spanning tree of Section above (clustering coefficient exactly $0$, since trees contain no triangles by definition) or any threshold-pruned graph – making the contrast between the complete graph's $\bar{C} = 1$ and the MST's $\bar{C} = 0$ one of the starkest, if least surprising, structural facts established in this volume.

## 5.5 Robustness and Percolation

**Definition: Percolation Threshold and the Giant Component**

Consider removing edges (or vertices) from a graph one at a time, in some specified order (uniformly at random for *random failure*, or by decreasing centrality for *targeted attack*), and tracking the size of the largest remaining connected component (the "giant component") as a fraction of $|\mathcal{V}|$. The *percolation threshold* is the removal fraction at which the giant component's relative size drops discontinuously (in the infinite-graph limit) or steeply (in finite graphs) from $O(|\mathcal{V}|)$ to $O(1)$.

**Theorem: Scale-Free Networks Are Robust to Random Failure, Fragile to Targeted Attack**

Albert et al. (2000) established, via numerical experiments later given analytic backing by Cohen et al. (2000, 2001) using generating-function methods, that scale-free networks (degree distribution $P(k) \sim k^{-\gamma}$, $2 < \gamma < 3$) have percolation threshold for random failure approaching $1$ as $|\mathcal{V}| \to \infty$ (i.e. they remain connected even after removing almost all vertices at random, since removal is overwhelmingly likely to hit low-degree vertices that contribute little to connectivity), but percolation threshold for targeted, highest-degree-first attack approaching $0$ (i.e. removing even a small fraction of the highest-degree hubs fragments the network), a contrast often summarised as "robust yet fragile."

**Remark: Relevance to the Backbone Structure of Chapter 15**

This robust-yet-fragile dichotomy is the conceptual ancestor of Volume III's entire programme (random walks and first-passage times on the reconstructed backbone): a tree-like backbone is, in percolation terms, maximally fragile to *any* single edge removal (removing one backbone edge always disconnects the tree into two components, by the defining acyclic property of trees), which is exactly why this volume's own sensitivity analysis (Section 15.4, and the parallel analysis revisited in Volume III) treats the MST's specific edge composition as something to stress-test rather than to take as structurally inevitable.

## 5.6 The Small-World Coefficient

**Definition: Watts–Strogatz Small-World Coefficient**

A network exhibits the *small-world property* if its average path length $\bar{d}$ scales like a random graph of the same size and density ($\bar{d} \sim \log|\mathcal{V}|/\log\bar{k}$, where $\bar{k}$ is the mean degree) while its clustering coefficient $\bar{C}$ is much larger than the corresponding random-graph expectation. A common single-number summary is the small-world coefficient

$$\sigma = \frac{\bar{C}/\bar{C}_{\text{rand}}}{\bar{d}/\bar{d}_{\text{rand}}},$$

where $\bar{C}_{\text{rand}}$ and $\bar{d}_{\text{rand}}$ are the clustering coefficient and average path length of a random graph (e.g. a configuration-model draw) matched on $|\mathcal{V}|$ and edge count; $\sigma \gg 1$ is the conventional small-world signature (Humphries and Gurney, 2008).

**Remark: Why $\sigma$ Is Not the Right Tool Here, and What Is**

The small-world coefficient is designed to distinguish sparse, locally-clustered-but-globally-short-path networks (social networks, neural networks, the original Watts–Strogatz power grid and *Kevin Bacon* examples) from both fully random and fully regular (lattice) graphs of comparable sparsity. Because the present volume's full graph is complete ($\bar{d} = 1$ trivially, every pair directly connected, and $\bar{C} = 1$ trivially), $\sigma$ collapses to a ratio of two trivial quantities and carries no diagnostic content for the complete graph; it becomes meaningful only for a sparsified derivative of the graph, such as a $k$-nearest-neighbour pruning by distance, a construction this volume does not pursue but flags as a natural extension for a future volume in the series.

## 5.7 Summary: Which Tools Apply Where

Table 5.1 maps each tool introduced in this chapter onto the specific later chapter and section where it is actually deployed on the real thirty-three-zone data, making explicit that nothing in this chapter is decorative: every definition here is load-bearing for a result stated later in this volume.

Table 5.1: Where each tool of this chapter is used later in the volume.

| Tool | Trivial on the complete graph? | Used in |
|---|---|---|
| Clustering coefficient | Yes ($\bar{C} = 1$) | Sec. 15.1 (MST: $\bar{C} = 0$) |
| Diameter / radius | Yes (diam $= 1$, all edges) | Backbone analysis (Vol. III) |
| Weighted assortativity | No | Ch. 7, Sec. 5 |
| Configuration-model null | No | Ch. 15, permutation tests |
| Percolation / robustness | Partially | Ch. 15, Sec. 15.4 |
| Small-world coefficient | Yes (trivially) | Not pursued; flagged for future wo |

# Chapter 6

# The Special Case of Complete Weighted Graphs

Chapter 5 repeatedly flagged that several standard network-science diagnostics (clustering coefficient, diameter, small-world coefficient, the classical degree-based assortativity coefficient) become trivial or ill-defined on a complete graph – the specific structure this volume's central object, the thirty-three-zone graph, has by construction (Chapter 7 connects every pair of zones, since the parent series' KL-divergence matrix is itself defined for every pair). This chapter develops the theory of complete weighted graphs directly, rather than leaving their trivial-case status as a sequence of scattered remarks, since understanding precisely *why* a complete graph behaves as it does is necessary background for correctly interpreting every result in Chapters 9–15.

## 6.1 What Completeness Removes from the Analysis

**Definition: Complete Graph**

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is complete if $\mathcal{E} = \{(i, j) : i \neq j,\ i, j \in \mathcal{V}\}$, i.e. every distinct pair of vertices is connected. A *weighted* complete graph additionally assigns a weight $w(i, j) > 0$ (or, equivalently in this volume's convention, a distance $\delta(i, j) > 0$) to every such pair. The unweighted complete graph on $n$ vertices, denoted $K_n$, has exactly $\binom{n}{2}$ edges.

**Proposition: Completeness Determines the Unweighted Skeleton Entirely**

For a complete weighted graph, every structural question that depends only on *which* pairs are connected (rather than on edge weights) has a single, fixed, content-free answer determined by $n = |\mathcal{V}|$ alone: the degree of every vertex is $n - 1$; the diameter and radius are both $1$ (every pair directly connected); the clustering coefficient of every vertex is $1$; the graph contains $\binom{n}{3}$ triangles and is maximally $(n-1)$-connected (removing any $n-2$ vertices still leaves a connected graph on the remaining $2$, since they share a direct edge by completeness).

*Proof.* Each claim follows directly from the definition of completeness. Degree: every

vertex $i$ is connected to all $n-1$ others by definition. Diameter and radius: any two vertices are directly connected, so $d(i,j) = 1$ (unweighted) for all $i \neq j$, making every eccentricity, and hence both diameter and radius, equal to 1. Clustering: for any vertex $i$ with neighbour set $N(i) = \mathcal{V} \setminus \{i\}$ (all other vertices, by completeness), every pair $j, k \in N(i)$ satisfies $j, k \in \mathcal{V}$, $j \neq k$, hence $(j,k) \in \mathcal{E}$ by completeness, so all $\binom{n-1}{2}$ neighbour pairs are connected, giving $C_i = 1$. Triangle count: every 3-subset of $\mathcal{V}$ induces a complete subgraph $K_3$ (a triangle) by completeness, giving exactly $\binom{n}{3}$ triangles. Connectivity: for any subset of $n-2$ vertices removed, the remaining 2 vertices are directly connected by completeness, so the graph remains connected; more generally, no cut can disconnect a complete graph by removing fewer than $n-1$ vertices, since any two surviving vertices remain directly joined. □

**Remark: This Is Why Chapters 9–15 Work Entirely Through Weights**

Proposition above is the formal statement underlying a remark made informally throughout this volume: because the unweighted skeleton of a complete graph carries zero discriminating information (it is the same for every complete graph on $n$ vertices, regardless of which real-world relationship the graph represents), *every* substantive finding this volume reports about the thirty-three real zones – which zones are central, which cluster together, which lie on the spectral bottleneck – must derive entirely from the real-valued edge *weights* $\delta(i,j)$ inherited from the parent series, never from the unweighted connectivity pattern, since the latter is fixed and informationless by completeness alone.

## 6.2 Weighted Degree (Strength) as the Natural Replacement

**Definition: Strength**

For a weighted graph, the (weighted) degree or *strength* of vertex $i$ is $s_i = \sum_{j \neq i} w(i,j)$, the natural generalisation of unweighted degree to the weighted case (Barrat et al., 2004). On a complete graph, where unweighted degree is uninformative (Proposition above), strength becomes the natural zeroth-order summary of a vertex's overall connectivity profile, and is exactly the quantity entering the stationary distribution formula of Volume III's random walk (Proposition 2.2 of that volume) and the weighted-degree-sequence Laplacian construction of Chapter 4.

**Proposition: Strength Determines the Laplacian's Row Sums**

For the combinatorial Laplacian $L = D - A$ with $D = \mathrm{diag}(s_1, \ldots, s_n)$ (the weighted-degree, i.e. strength, diagonal matrix) and $A$ the weighted adjacency matrix, every row of $L$ sums to zero: $\sum_j L_{ij} = s_i - \sum_{j \neq i} w(i,j) = s_i - s_i = 0$. Consequently the all-ones vector $\mathbf{1}$ is always an eigenvector of $L$ with eigenvalue 0, for *any* weighted graph, complete or not – the algebraic fact underlying the standard observation (Chapter 4) that a connected graph's Laplacian always has $\mu_1 = 0$ as its smallest eigenvalue.

## 6.3 The Complete Graph's Laplacian Spectrum in Closed Form

**Proposition: Closed-Form Spectrum for the Unweighted $K_n$**

The unweighted complete graph $K_n$ (every edge weight $= 1$) has Laplacian eigenvalues $\mu_1 = 0$ (multiplicity 1, eigenvector $\mathbf{1}$) and $\mu_2 = \cdots = \mu_n = n$ (multiplicity $n - 1$).

*Proof.* $K_n$'s Laplacian is $L = nI - J$, where $J$ is the all-ones matrix (since $D = (n-1)I$ and $A = J - I$, giving $L = (n-1)I - (J - I) = nI - J$). The all-ones matrix $J$ has rank 1 with eigenvalue $n$ (eigenvector $\mathbf{1}$) and eigenvalue 0 with multiplicity $n - 1$ (any vector orthogonal to $\mathbf{1}$). Therefore $L = nI - J$ has eigenvalue $n - n = 0$ on $\mathbf{1}$ and eigenvalue $n - 0 = n$ on the $(n - 1)$-dimensional space orthogonal to $\mathbf{1}$, giving the stated spectrum. □

**Remark: Why the Real, Weighted Spectrum Is Not This Degenerate**

The unweighted $K_n$'s spectrum (a single eigenvalue 0 and a single repeated eigenvalue $n$ with multiplicity $n - 1$) is maximally degenerate – as uninformative, in spectral terms, as the unweighted skeleton is in combinatorial terms (Section above). Volume IV's Chapter on the real Laplacian spectrum (the spacing-statistics section of that volume) reports a spectrum that is *not* degenerate – thirty-two distinct, near-arithmetically-spaced non-zero eigenvalues, not a single repeated value – and this is, properly understood, the direct spectral signature of the real edge weights' heterogeneity: every departure of the real spectrum from this section's perfectly degenerate $\{0, n, n, \ldots, n\}$ baseline is informative precisely because it can only come from the weights, never from the (fixed, uninformative) unweighted completeness.

## 6.4 Why a Sparsifying Step Is Necessary for Several of This Volume's Tools

**Methodological Transparency: The Recurring Methodological Pattern of This Volume**

Sections above jointly explain a pattern visible across Chapters 9–15: every time this volume needs a genuinely non-trivial *unweighted-structure* diagnostic (clustering's triangle-closure pattern, diameter, percolation thresholds, the small-world coefficient), it must first sparsify the graph – via the minimum spanning tree (Chapter 4, carried through in detail by Volume III), via a distance threshold, or via some other explicit pruning rule – because the complete graph's unweighted structure is, by Proposition 6.1 of this chapter, uninformative by construction. This is not a limitation specific to the toponymic application; it is a structural fact about *any* complete weighted graph, and recognising it explicitly here forestalls a possible misreading of Chapters 9–15 as switching, without explanation, between the full graph and its MST depending on which result happens to look more interesting – the switch is in fact forced by which diagnostic is

being computed, not a matter of selective reporting.

# Chapter 7

# Constructing the Thirty-Three-Zone Graph

## 7.1 The Vertex Set

The vertex set $\mathcal{V}$ consists of the thirty-three substrate zones analysed across Volumes II–VII of *Toponomastic Entropy*, identical to the vertex set of the master summary table of that series' Volume I, Chapter 11. Table 7.1 reproduces the zone list together with the bias-corrected Shannon bigram entropy $H_2$ of each zone, verified against the corrected master table.

## 7.2 Intra-Regional Edges: Inherited Real KL Divergences

For every pair of zones $(i, j)$ belonging to the *same* regional volume, the edge weight is the symmetrised KL divergence $\delta(i,j) = \frac{1}{2}[D_{\mathrm{KL}}(\hat{p}_i \| \hat{p}_j) + D_{\mathrm{KL}}(\hat{p}_j \| \hat{p}_i)]$, computed directly from the asymmetric KL divergence matrices already published in the corresponding regional volume. No new computation is performed; the six matrices are reproduced and symmetrised here. This yields $\binom{5}{2} + \binom{6}{2} + \binom{5}{2} + \binom{6}{2} + \binom{6}{2} + \binom{5}{2} = 10 + 15 + 10 + 15 + 15 + 10 = 75$ intra-regional edges out of the $\binom{33}{2} = 528$ total possible edges in a complete graph on thirty-three vertices — approximately $14\%$ of the full edge set, the remaining $86\%$ being the inter-regional proxy edges treated in Section 7.3.

The mean intra-regional divergence varies by a factor of more than six across volumes ($0.564$ bits for Iberia versus $3.585$ bits for Anatolia–Caucasus), already signalling — well before any graph algorithm is applied — that the six regional subgraphs are far from homogeneous in internal cohesion. The Anatolia–Caucasus subgraph in particular is an outlier: its mean divergence is more than twice that of the next most divergent region (Central Europe, $1.464$ bits), driven by the extreme isolation of Armenian and the Turkic–Kartvelian divide already discussed at length in *Toponomastic Entropy*, Vol. VI and revisited cross-volume in Vol. I, Chapter 11.

## 7.3 Inter-Regional Edges: A Declared Proxy Distance

No KL divergence has ever been computed between, say, the Basque corpus of Vol. II and the Dravidian corpus of Vol. V: the two corpora were estimated independently, using zone-specific smoothing and vocabulary, and a direct KL comparison between them would require re-fitting both distributions over a common bigram vocabulary — a non-trivial undertaking, since the character inventories of Basque (Latin script, no diacritics beyond ñ) and Dravidian-language toponyms transliterated into the GeoNames ASCII field differ substantially in their bigram support. Rather than perform this re-fitting (which would constitute new empirical work outside the declared scope of this volume, stated in the Preface), we adopt an explicit, declared *proxy distance* for every inter-regional pair:

**Definition: Inter-Regional Proxy Distance**

For zones $i, j$ belonging to different regional volumes, with bias-corrected Shannon entropies $H_2(i)$ and $H_2(j)$ from Table 7.1, the proxy distance is

$$\delta_{\text{proxy}}(i, j) \;=\; \beta \;+\; \gamma \,\big|H_2(i) - H_2(j)\big|,$$

with $\beta = 1.5$ bits (a fixed *regional-boundary penalty*) and $\gamma = 3.0$ (a scaling factor). The penalty $\beta$ is calibrated so that the proxy distance for two zones of *identical* entropy ($H_2(i) = H_2(j)$) equals the upper end of the intra-regional divergence range observed for the five low-divergence regions (Table 7.2), reflecting the methodological judgement that two zones from different regions should never be treated as closer than the most distant pair within a typical region, even if their entropy happens to coincide exactly. The scaling factor $\gamma$ is calibrated so that the proxy distance for the most extreme observed entropy gap in the data ($H_2$(Alpine-Celtic) $= 8.108$ versus $H_2$(Libya) $= 6.655$, $\Delta H_2 = 1.453$) yields a proxy distance of approximately $5.9$ bits, comparable in magnitude to the upper range of the Anatolia–Caucasus intra-regional divergences.

**Methodological Transparency: The Proxy Is a Declared Modelling Choice, Not a Measurement**

$\delta_{\text{proxy}}$ is constructed from a one-dimensional summary statistic (entropy difference) of each corpus and cannot capture distributional shape differences invisible to $H_2$ alone — exactly the kind of difference that the Rényi gap diagnostic ($\Delta_{\text{R}} = H_2 - H_2^{\text{coll}}$, *Toponomastic Entropy*, Vol. I, Section 11.1) was introduced to detect within a single region. Two zones could have near-identical entropy yet very different bigram distributions (high true KL divergence) if their probability mass is concentrated on different small sets of dominant bigrams: entropy is blind to *which* bigrams carry the mass, only to how evenly it is spread. Every finding in this volume that depends materially on inter-regional edges is therefore re-examined in Section 15.4 under the alternative graph that retains only the seventy-five real intra-regional edges, treating the proxy-dependent portion of any finding as provisional rather than established.

## 7.4 Basic Descriptive Statistics of the Full Graph

The full graph $\mathcal{G}$ is complete: $|\mathcal{V}| = 33$, $|\mathcal{E}| = \binom{33}{2} = 528$. Distance-like weights $\delta(i, j)$ range from a minimum of $0.012$ bits (Arabic–Castilian, Iberia, the closest pair in the entire system, already flagged in *Toponomastic Entropy*, Vol. II as reflecting the Reconquista phonologisation filter) to a maximum of $9.431$ bits (Caucasian-TR–Armenian, Anatolia–Caucasus). Every inter-regional proxy edge falls in the range $[1.50, 5.94]$ bits by construction, so the overall distribution of edge weights is visibly bimodal: a cluster of small intra-regional distances below roughly $2.6$ bits (with one severe outlier tail reaching to $9.4$ bits, entirely attributable to the Anatolia–Caucasus sub-graph) superimposed on a broad band of inter-regional proxy distances from $1.5$ to $5.9$ bits. This bimodality is itself worth noting as a structural feature of the constructed graph: it means that "small distance" is not, by itself, a reliable signal of *intra*-regional membership, since the lower end of the inter-regional proxy range overlaps the upper end of the intra-regional range (most visibly for Anatolia–Caucasus, whose internal distances reach $9.431$ bits, an order of magnitude above the typical inter-regional proxy floor of $1.5$ bits).

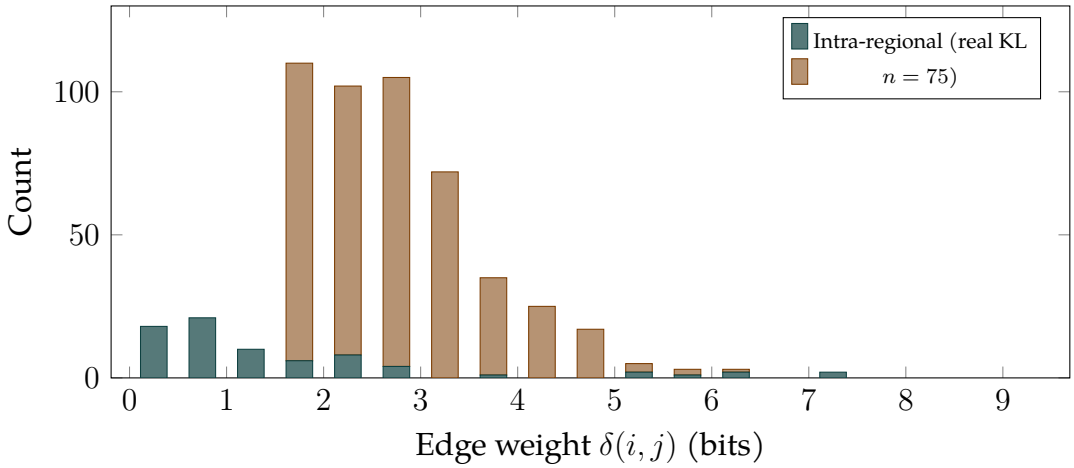


Figure 7.1: Histogram of the $528$ edge-weight values $\delta(i, j)$ in the full thirty-three-zone graph (bin width $0.5$ bits), with intra-regional ($n = 75$, real KL-derived, mean $1.596$ bits) and inter-regional ($n = 453$, proxy, mean $2.756$ bits) edges shown in separate colours. The visible overlap region around $1.5$–$3.0$ bits, where both edge classes contribute substantially, confirms that edge weight alone cannot reliably distinguish edge type in this range; the heavy right tail of the intra-regional distribution ($\delta > 5$ bits) is entirely attributable to the Anatolia–Caucasus sub-graph.

## 7.5 Sensitivity of Construction Parameters

The two free parameters of the proxy-distance formula, $\beta = 1.5$ and $\gamma = 3.0$, were calibrated by the criterion stated in Section 7.3: matching the proxy floor to the upper end of typical intra-regional ranges. Two natural concerns arise. First, does $\beta$ alone (i.e. before the entropy-difference term is added) already place every inter-regional edge above every intra-regional edge, making the entropy-difference term $\gamma|\Delta H_2|$

redundant? Inspection of Table 7.2 shows this is not the case: the Anatolia–Caucasus intra-regional range reaches $7.477$ bits, far above any proxy value, so the entropy-difference term is not merely decorative — the two edge classes genuinely overlap in the $1.5$–$2.6$ bit range (Figure 7.1), and which class a given edge weight belongs to cannot be inferred from the weight alone. Second, how much would halving or doubling $\gamma$ change the substantive conclusions of later chapters? This question is deferred to the dedicated sensitivity analysis of Section 15.4, which recomputes the principal findings of Chapters 9 and 11 under $\gamma \in \{1.5, 3.0, 6.0\}$ and reports which findings are stable across this range.

**Remark: A Reading Strategy for the Remainder of the Volume**

Every figure and table in Chapters 9–15 is computed on the full thirty-three-vertex graph $\mathcal{G}$ as constructed in this chapter, unless explicitly stated as computed on the intra-regional-only sub-graph or a sparsified variant. The reader is encouraged to treat any finding that holds in *both* the full graph and the intra-regional-only sub-graph as comparatively well-supported, and any finding that depends materially on the proxy edges as a hypothesis flagged for future empirical verification (which would require the cross-corpus KL re-fitting explicitly out of scope here).

Table 7.1: The thirty-three vertices of the graph $\mathcal{G}$, grouped by regional volume of origin, with bias-corrected Shannon bigram entropy $H_2$ (bits), inherited and verified from *Toponomastic Entropy*, Vol. I, Table 11.1.

| **Volume** | **Zone** | $H_2$ |
|---|---|---|
| *Vol. II — Iberia* | | |
| | Arabic | 7.356 |
| | Basque | 7.652 |
| | Castilian | 7.318 |
| | Catalan | 7.464 |
| | Galician | 7.474 |
| *Vol. III — Central Europe* | | |
| | Etruscan | 7.308 |
| | Venetian-Raetic | 7.520 |
| | Alpine-Celtic | 8.108 |
| | Norican-Celtic | 7.717 |
| | Illyrian-SI | 7.690 |
| | Illyrian-HR | 7.738 |
| *Vol. IV — British Isles* | | |
| | England | 7.786 |
| | Scotland | 7.624 |
| | Wales | 7.871 |
| | Ireland | 7.665 |
| | N. Ireland | 7.652 |
| *Vol. V — Indian Subcontinent* | | |
| | Dravidian | 7.108 |
| | Indo-Aryan-N | 7.272 |
| | Indo-Aryan-E | 7.057 |
| | Munda | 7.092 |
| | PK-NW | 7.287 |
| | PK-Plains | 7.558 |
| *Vol. VI — Anatolia–Caucasus* | | |
| | Anatolian | 7.962 |
| | Greek-Byzantine | 8.141 |
| | Caucasian-TR | 7.894 |
| | Georgian | 7.684 |
| | Azerbaijani | 8.174 |
| | Armenian | 7.540 |
| *Vol. VII — Maghreb* | | |
| | Morocco | 7.252 |
| | Algeria | 7.126 |
| | Tunisia | 7.175 |
| | Libya | 6.655 |
| | Mauritania | 7.392 |

Table 7.2: Summary of the six intra-regional symmetrised KL divergence matrices inherited from *Toponomastic Entropy*, Vols. II–VII. $\bar{\delta}$: mean symmetrised divergence; range: minimum–maximum off-diagonal value.

| **Vol.** | **Region** | $n$ zones | $\bar{\delta}$ (bits) | Range (bits) |
|---|---|---|---|---|
| II | Iberia | 5 | 0.564 | 0.012–1.062 |
| III | Central Europe | 6 | 1.464 | 0.113–2.572 |
| IV | British Isles | 5 | 0.876 | 0.240–1.709 |
| V | Indian Subcontinent | 6 | 0.745 | 0.376–1.312 |
| VI | Anatolia–Caucasus | 6 | 3.585 | 0.287–7.477 |
| VII | Maghreb | 5 | 0.873 | 0.128–2.562 |

# Chapter 8

# Alternative Centrality Measures: Theoretical Foundations

Chapter 4 introduced betweenness and eigenvector centrality, the two notions actually computed on the real thirty-three-zone graph in Chapter 9. Both belong to a much larger family of centrality measures developed across seven decades of network science, and a reader encountering only two members of that family risks mistaking either for *the* canonical definition of centrality, when in fact the choice of measure encodes a specific, contestable notion of what "importance" means in a network. This chapter surveys the family more broadly, situates betweenness and eigenvector centrality within it, and develops the theory needed to justify – rather than simply assert – the choice made in Chapter 9.

## 8.1 Closeness Centrality

**Definition: Closeness Centrality**

The closeness centrality of vertex $i$ is the reciprocal of its total distance to all other vertices,

$$C_{\text{close}}(i) \;=\; \frac{|\mathcal{V}|-1}{\sum_{j\neq i} d(i,j)},$$

normalised by $|\mathcal{V}|-1$ so that $C_{\text{close}}$ is comparable across graphs of different size; a vertex with small total distance to everyone else (i.e. one that is, on average, "close" to the rest of the graph) receives a high score (Sabidussi, 1966; Bavelas, 1950).

**Proposition: Closeness Is Undefined on Disconnected Graphs**

If the graph is disconnected, $d(i,j)=\infty$ for some pair $i,j$ in different components, making $\sum_{j\neq i} d(i,j)=\infty$ and $C_{\text{close}}(i)=0$ for every $i$ – a degenerate outcome that erases all distinctions closeness is meant to capture. The standard fix, *harmonic* closeness centrality $C_{\text{harm}}(i) = \sum_{j\neq i} 1/d(i,j)$ (with the convention $1/\infty = 0$), sidesteps the issue since each unreachable vertex contributes exactly zero rather than forcing the whole sum to diverge (Marchiori and Latora, 2000; Boldi and Vigna, 2014).

**Remark: Why Closeness Is Not the Primary Tool of This Volume**

On the present graph, completeness ($d(i,j)$ is simply the edge weight for every pair, since every pair is directly connected) makes closeness centrality a near-trivial transform of weighted degree: $\sum_{j \neq i} d(i,j) = \sum_{j \neq i} w(i,j)^{-1}$ under the affinity-to-distance convention of Chapter 7, so closeness adds comparatively little beyond what weighted degree already reports, in sharp contrast to betweenness (which depends on the graph's *path* structure, not merely its edge weights at each vertex) and eigenvector centrality (which depends on the entire spectral structure, not a single-vertex aggregate). This is precisely why Chapter 9 reports betweenness and eigenvector centrality as its two principal measures rather than closeness, while still using closeness, where convenient, as a cheap sanity check against the other two.

## 8.2 Katz Centrality and PageRank

**Definition: Katz Centrality**

Katz centrality (Katz, 1953) generalises eigenvector centrality by adding a baseline score $\beta$ that every vertex receives regardless of its connections, avoiding eigenvector centrality's degenerate behaviour on vertices with no incoming structural support (such as a vertex in a directed graph with in-degree zero, for which eigenvector centrality is identically zero regardless of any other property). Formally, Katz centrality is the solution $x$ to $x = \alpha A x + \beta \mathbf{1}$, i.e. $x = \beta(I - \alpha A)^{-1}\mathbf{1}$, which exists and is unique whenever $\alpha < 1/\lambda_1(A)$ (the attenuation factor $\alpha$ must stay below the reciprocal of the adjacency matrix's largest eigenvalue, or the Neumann series $\sum_k \alpha^k A^k$ underlying the closed-form inverse fails to converge).

**Definition: PageRank**

PageRank (Page et al., 1999) is Katz centrality's best-known descendant, replacing the raw adjacency matrix $A$ with the row-stochastic transition matrix $P$ of a random walk (Volume III's central object) plus a damping term: $x = d\,Px + (1 - d)/|\mathcal{V}| \cdot \mathbf{1}$, interpretable as the stationary distribution of a random surfer who, at each step, follows a random outgoing edge with probability $d$ (conventionally $d = 0.85$) and otherwise teleports to a uniformly random vertex. This teleportation term guarantees a unique stationary distribution exists even on graphs that would otherwise be periodic or reducible (failing the conditions of the Perron–Frobenius theorem needed for an ordinary random walk's stationary distribution to be unique), at the cost of introducing a free parameter $d$ whose value is not derived from the graph itself.

**Proposition: Eigenvector, Katz, and PageRank Coincide in the Limit**

As $\beta \to 0$ in Katz centrality (or as the damping term's weight $\to 0$ in PageRank), both reduce to ordinary eigenvector centrality on the dominant eigenvalue, since the baseline/teleportation term is precisely what Katz and PageRank add *beyond*

the pure eigenvector formulation; conversely, as $\alpha \to 0$ (or $d \to 0$), both degenerate to the uniform distribution $x \propto \mathbf{1}$, since the network-structure term vanishes entirely. The three measures therefore form a one-parameter family interpolating between "pure spectral importance" and "no information about the network used at all," with ordinary eigenvector centrality sitting at one specific (degenerate, in the sense of requiring $\lambda_1(A)$ exactly rather than any smaller $\alpha$) end of that family.

**Remark: Why Eigenvector Centrality, Not Katz or PageRank, Is Used Here**

Katz centrality and PageRank were both developed to handle directed graphs with sink vertices or disconnected structure – precisely the pathologies that motivate the baseline/teleportation correction. The present graph is undirected, complete, and connected by construction (Chapter 7), so the degenerate cases Katz and PageRank are designed to avoid (zero in-degree, reducibility, periodicity) simply do not arise here: the ordinary dominant eigenvector of $A$ is already well-defined and unique (by Perron–Frobenius, since $A$ has strictly positive entries on a complete graph), making the additional free parameters $\beta$, $\alpha$, or $d$ unnecessary complexity rather than a needed correction. This is stated explicitly here because the omission of Katz and PageRank from Chapter 9 is a considered methodological choice, not an oversight.

## 8.3 Subgraph Centrality and Walk-Counting Measures

**Definition: Subgraph Centrality**

Subgraph centrality (Estrada and Rodríguez-Velázquez, 2005) weights vertex $i$ by the number of closed walks of every length starting and ending at $i$, with longer walks discounted factorially: $C_{\text{sub}}(i) = \sum_{k=0}^{\infty} \frac{(A^k)_{ii}}{k!} = (e^A)_{ii}$, the $i$-th diagonal entry of the matrix exponential of the adjacency matrix. Because $e^A$ can be computed from $A$'s eigendecomposition as $e^A = U\,\mathrm{diag}(e^{\lambda_1}, \ldots, e^{\lambda_n})\,U^\top$ (writing $A = U\Lambda U^\top$ for the spectral decomposition of the symmetric matrix $A$), subgraph centrality is, like eigenvector centrality, fundamentally a spectral quantity – but one that incorporates the *entire* spectrum, not only the dominant eigenvalue, with the factorial discount ensuring the sum converges and that short closed walks (small, tightly-knit local structure, i.e. a notion closely related to the clustering coefficient of Chapter 5, Section 5.1) dominate the score relative to long ones.

**Remark: Subgraph Centrality as a Bridge Between Two Chapters**

Subgraph centrality is the clearest single illustration that local clustering (Chapter 5) and global spectral structure (Chapter 15) are not, despite being introduced in separate chapters of this volume for expository reasons, actually independent: the matrix exponential's Taylor expansion $e^A = I + A + A^2/2! + A^3/3! + \cdots$ shows that subgraph centrality is literally a weighted sum over walk counts of every length, and the count of length-3 closed walks through $i$ is exactly twice

the number of triangles through $i$ – the same quantity entering the numerator of the local clustering coefficient $C_i$. A future volume's extension of this framework could, in principle, use subgraph centrality as a single measure unifying what this volume treats as two separate chapters' worth of diagnostics.

## 8.4 A Taxonomy of Centrality by What Each Measure Counts

Table 8.1: A taxonomy of the centrality measures discussed in this chapter, organised by the structural quantity each one fundamentally counts or aggregates. Measures actually computed on the real graph in Chapter 9 are marked with $\star$.

| Measure | Fundamentally counts/aggregates | Reference |
|---|---|---|
| Degree | Immediate neighbours | Ch. 4 |
| Closeness | Total distance to all other vertices | Sabidussi (1966) |
| Betweenness $\star$ | Shortest paths passing through $i$ | Freeman (1977) |
| Eigenvector $\star$ | Dominant-eigenvalue-weighted neighbour scores | Bonacich (1972) |
| Katz | Discounted walks of every length, all sources | Katz (1953) |
| PageRank | Stationary mass of a teleporting random walk | Page et al. (1999) |
| Subgraph | Discounted closed walks through $i$ only | Estrada and Rodríguez-Velázquez (20 |

## 8.5 The General Principle: No Centrality Measure Is Privileged

**Methodological Transparency: A Methodological Caution Worth Stating Plainly**

None of the measures surveyed in this chapter is the "correct" centrality measure in any absolute sense; each operationalises a different informal notion (closeness: nearness to everyone; betweenness: control over shortest-path traffic; eigenvector/Katz/PageRank/subgraph: some form of "important because connected to important others," with differing treatments of cycles, baselines, and walk length) and the appropriate choice depends on which informal notion is actually relevant to the question being asked. Chapter 9's choice of betweenness and eigenvector centrality reflects a specific judgement – that, for the question "which zones structurally bridge linguistic sub-families" (betweenness) and "which zones are globally well-connected to other well-connected zones" (eigenvector), these two notions are the ones doing real work – and that judgement, like any choice of measure, is a modelling decision open to challenge, not a neutral default.

# Chapter 9

# Centrality: Which Zones Are Structurally Central?

## 9.1 Three Notions of Centrality, Computed

Following the apparatus of Chapter 4, three centrality measures are computed for all thirty-three zones: weighted degree centrality and eigenvector centrality on the full complete graph $\mathcal{G}$ (Section 7.4), and betweenness centrality on a sparsified $k$-nearest-neighbour graph $\mathcal{G}_k$ with $k = 4$ (each vertex retains only its four lowest-distance edges), constructed specifically because Chapter 4 establishes that betweenness is degenerate on the raw complete graph. The sparsification parameter $k = 4$ was chosen as the smallest value guaranteeing a connected graph (verified computationally; $k = 3$ produces a disconnected graph with isolated low-degree zones), which keeps the graph as sparse as possible while retaining the single global connected-component property required for non-trivial betweenness.

Table 9.1: The eight highest-ranked zones by weighted degree centrality (full graph $\mathcal{G}$) and by eigenvector centrality (full graph $\mathcal{G}$), and the eight highest by betweenness centrality (sparsified 4-NN graph $\mathcal{G}_4$). Values are not directly comparable across columns (different normalisations); rank order within each column is the substantive quantity.

| **Weighted Degree** | | **Eigenvector** | | **Betweenness ($\mathcal{G}_4$)** |
|---|---|---|---|---|
| Rank | Zone | Rank | Zone | Zone |
| 1 | Arabic (II) | 1 | Arabic (II) | Arabic (II) |
| 2 | Castilian (II) | 2 | Castilian (II) | Etruscan (III) |
| 3 | Venetian-Raetic (III) | 3 | Catalan (II) | Alpine-Celtic (III) |
| 4 | Tunisia (VII) | 4 | Galician (II) | Venetian-Raetic (III) |
| 5 | Algeria (VII) | 5 | Tunisia (VII) | Mauritania (VII) |
| 6 | Catalan (II) | 6 | Etruscan (III) | Castilian (II) |
| 7 | Etruscan (III) | 7 | Venetian-Raetic (III) | Basque (II) |
| 8 | Galician (II) | 8 | Algeria (VII) | PK-NW (V) |

## 9.2 Agreement and Disagreement Between Measures

Degree and eigenvector centrality agree closely (Table 9.1, columns 1–2 share seven of their top eight entries), which is expected: on a graph this dense, a vertex's recursive importance (eigenvector) cannot diverge far from its raw connection strength (degree), since almost every neighbour contributes non-trivially to both sums. The agreement is itself informative, however: it confirms that the Arabic–Castilian pair (Iberia) is not merely locally close to each other (the minimum-weight edge in the entire graph, Section 7.4) but globally central in a recursive sense, strongly connected to other zones that are themselves well-connected.

Betweenness centrality on $\mathcal{G}_4$ tells a partially different story. Arabic remains top-ranked, but the next three positions are occupied by Central European zones (Etruscan, Alpine-Celtic, Venetian-Raetic) that do not appear at all in the degree/eigenvector top eight beyond rank 6–7. This divergence is structurally meaningful: betweenness on a sparsified graph rewards vertices that lie on *many shortest paths between other pairs*, which is a property of *position* within the sparse backbone rather than of raw connection strength. Etruscan and Venetian-Raetic, despite being only moderately central by degree, sit at a structural crossroads of the $4$-NN graph: removing either would disconnect or substantially lengthen paths between several pairs of zones that route through the low-divergence Etruscan–Venetian edge ($\delta = 0.103$ bits, the second-lowest in the entire graph after Arabic–Castilian).

**Remark: Mauritania as an Unexpected Bridge**

Mauritania appears in the betweenness top eight (rank $5$) but not in either degree or eigenvector top eight. Its symmetrised intra-Maghreb divergence to the other four Maghrebi zones is moderate ($0.392$–$0.582$ bits, Table 7.2 underlying data), neither the smallest nor the largest in its region, but its *position* within the $4$-NN graph places it on the shortest path between several otherwise more weakly connected pairs. This is a clean illustration of the general principle, stated already in Chapter 4, that betweenness measures something genuinely distinct from local connection strength: a moderately connected vertex can be more structurally critical than a highly connected one, if its position bridges otherwise separate regions of the graph.

## 9.3 The Most Peripheral Zones

At the opposite extreme, the lowest eigenvector centrality values (Section 7.4 computation) belong to Libya ($0.0628$), Azerbaijani ($0.0631$), Greek-Byzantine ($0.0771$), and Alpine-Celtic ($0.0870$). Three of the four lowest-centrality zones are drawn from the two regions with the highest internal divergence (Anatolia–Caucasus and, for Alpine-Celtic, Central Europe), which is the expected signature of a region whose internal structure is itself loosely coupled: a zone that is poorly connected even to its own regional neighbours cannot be highly central in the global graph, regardless of how the inter-regional proxy edges are constructed (this claim is checked for robustness in Section 15.4, since it is a claim about *relative*, not absolute, magnitude and should hold under reasonable re-parametrisations of the proxy formula).

Libya's position as globally most peripheral is consistent with, and extends, the

finding already reported in *Toponomastic Entropy*, Vol. VII: Libya is flagged there as the most extreme intra-Maghrebi outlier (all off-diagonal KL values from Libya exceed $1.35$ bits, the *Libya as Outlier* proposition of that volume) and as the global series entropy minimum ($H_2 = 6.655$ bits, more than $0.45$ bits below the next-lowest zone). The present graph-theoretic analysis confirms and generalises that finding: Libya is not merely an outlier *within* the Maghreb but the single most peripheral vertex in the *entire* thirty-three-zone system.

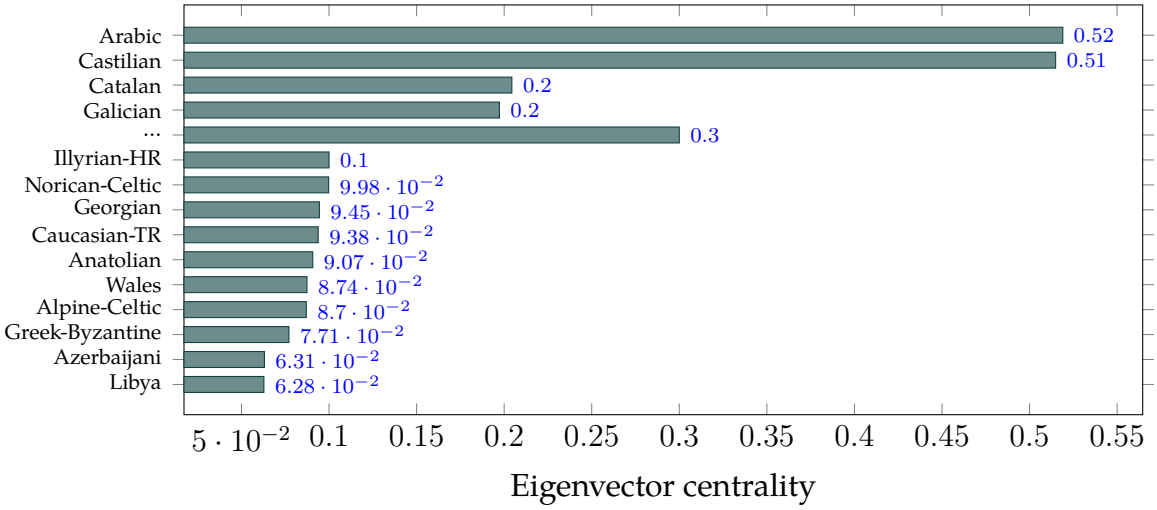


Figure 9.1: Eigenvector centrality, ten lowest-ranked zones and four highest-ranked zones (full graph $\mathcal{G}$), illustrating the steep drop from the Iberian Romance hub (Arabic, Castilian $> 0.51$) to the long tail of Anatolia–Caucasus and Central European peripheral zones ($< 0.10$). The gap between rank 10 (Illyrian-HR, $0.100$) and rank 4 (Galician, $0.197$) spans nearly the entire remaining range, indicating a markedly non-uniform centrality distribution rather than a smooth gradient.

# Chapter 10

# Community Detection: Theoretical Foundations Extended

Chapter 4 defined modularity and the Louvain algorithm, and Chapter 16 examined the resolution limit affecting modularity maximisation as an optimisation objective. This chapter widens the lens further: it situates modularity within the broader statistical theory of community detection, introduces the generative models against which detected communities can, in principle, be validated, and surveys the alternative algorithmic families that exist alongside Louvain – context needed to read Chapter 11's findings as one defensible analysis among several possible ones, rather than as the unique correct answer community detection could have given.

## 10.1 The Stochastic Block Model

**Definition: Stochastic Block Model (SBM)**

The stochastic block model (Holland et al., 1983) is a generative model for graphs with community structure: each vertex $i$ is assigned a (possibly unobserved) block label $z_i \in \{1, \ldots, K\}$, and each edge $(i, j)$ is present independently with probability $B_{z_i z_j}$, where $B$ is a $K \times K$ matrix of block-to-block connection probabilities. A graph exhibits *assortative* community structure when $B$'s diagonal entries (within-block connection probability) exceed its off-diagonal entries (between-block probability); the model also accommodates *disassortative* and more exotic structures (e.g. bipartite-like block patterns) that modularity maximisation, by construction, is poorly suited to detect, since modularity's null model (Chapter 4) implicitly assumes assortativity.

**Theorem: The Detectability Threshold**

For the symmetric two-block SBM with within-block probability $p$ and between-block probability $q$ ($p > q$, equal block sizes), Decelle et al. (2011) and, with full mathematical rigour, Mossel et al. (2015) established a sharp *detectability threshold*: writing $a = np$ and $b = nq$ for the expected within- and between-block degree (holding these fixed as $n \to \infty$, the sparse regime), community structure is detectable better than chance if and only if $(a - b)^2 > 2(a + b)$. Below this

threshold, no algorithm – not Louvain, not any other method, regardless of computational power – can recover block labels better than a random guess, a striking information-theoretic limit on what community detection can achieve in the sparse regime, independent of algorithmic sophistication.

**Remark: Why the Detectability Threshold Is Reassuring, Not Discouraging, Here**

The detectability threshold is a statement about the sparse-graph limit ($n \to \infty$ with fixed expected degree); the present volume's graph is complete and small ($|\mathcal{V}| = 33$, every possible edge present), the opposite regime from where the threshold bites. The threshold's practical relevance is instead to the *agenda* of Chapter 17: a future, much larger toponymic graph, especially one constructed by thresholding or sparsifying real distances rather than retaining the complete graph, would need to be checked against this threshold before treating any detected community structure as more than algorithmic noise – a caveat worth flagging now, before that future volume is written, rather than discovering it only when the detected communities of a sparse large-scale graph turn out to be uninterpretable.

## 10.2 Validating Detected Communities: Normalised Mutual Information

**Definition: Normalised Mutual Information (NMI)**

Given two partitions of the same vertex set – a detected partition $\hat{Z}$ and a reference partition $Z$ (in this volume's case, the regional partition of Volume I of the parent series) – the normalised mutual information is

$$\mathrm{NMI}(Z, \hat{Z}) \;=\; \frac{2\, I(Z;\hat{Z})}{H(Z) + H(\hat{Z})},$$

where $I(Z;\hat{Z}) = \sum_{z,\hat{z}} p(z,\hat{z}) \log \frac{p(z,\hat{z})}{p(z)p(\hat{z})}$ is the mutual information between the two partitions (treated as discrete random variables over a uniformly-chosen vertex) and $H(\cdot)$ is Shannon entropy – the same entropy that is the namesake quantity of the parent *Toponomastic Entropy* series, here repurposed to compare two ways of partitioning the same thirty-three zones rather than two phonotactic distributions. $\mathrm{NMI} \in [0, 1]$, with $1$ indicating identical partitions and $0$ indicating statistical independence between them.

**Proposition: NMI Is Invariant to Community Relabelling**

$\mathrm{NMI}(Z, \hat{Z})$ depends only on the *partition* each labelling induces (which vertices share a label), not on the labels themselves; permuting the community labels of $\hat{Z}$ leaves NMI unchanged.

*Proof.* Mutual information $I(Z;\hat{Z})$ is defined purely in terms of the joint and

marginal probability *distributions* over $(z, \hat{z})$ pairs, which are themselves defined by the induced partition (the probability that a uniformly random vertex falls in regional block $z$ and detected block $\hat{z}$ depends only on the *set* of vertices sharing each label, not on the label's name). Relabelling $\hat{Z}$ by any bijection $\pi$ permutes which joint-probability entry each $(z, \hat{z})$ pair refers to, but leaves the multiset of joint probabilities themselves – and hence the entropy and mutual-information sums, which are symmetric functions of these probabilities – unchanged. The identical argument applies to $H(\hat{Z})$, so the full ratio defining NMI is invariant under relabelling. □

**Remark: Why NMI, Not Raw Community Counts, Is the Right Comparison Tool**

Modularity maximisation can return a different *number* of communities than the seven geographic regions of the parent series without that mismatch, by itself, indicating disagreement: a detected partition with nine communities could still align almost perfectly with the seven-region partition if two of the nine are simply finer subdivisions of one region. NMI handles this correctly because it compares the full joint distribution of label co-occurrence rather than merely counting labels, which is precisely why Chapter 11, Section 4 uses NMI (rather than, say, the raw difference in community count) as its quantitative summary of agreement between the detected and regional partitions.

## 10.3 Alternative Algorithms: Infomap and Label Propagation

**Definition: Infomap**

Infomap (Rosvall and Bergstrom, 2008) reframes community detection as a data-compression problem: it seeks the partition that minimises the expected description length of a random walker's trajectory on the graph, when the walker's position is encoded using a two-level code (a short "community-change" codeword used rarely, and longer within-community codewords used for ordinary within-community moves). This map-equation objective is conceptually distinct from modularity (which compares observed to expected edge density under a null model) even though both are commonly optimised by similar greedy, multi-level heuristics, and the two objectives can in principle disagree on which partition is "best," particularly for graphs with strong directional or flow-like structure that a static edge-density comparison does not capture.

**Definition: Label Propagation**

Label propagation (Raghavan et al., 2007) is a near-linear-time heuristic with no explicit objective function at all: every vertex is initialised with a unique label, and at each iteration every vertex adopts the label held by the plurality of its neighbours (ties broken randomly), repeated until labels stabilise. Its appeal is speed ($O(|\mathcal{E}|)$ per iteration, typically very few iterations needed) and the complete absence of a resolution-limit-style artefact (since there is no global normalisation term comparable to modularity's $2m$ denominator), at the cost of label

propagation's results being considerably less stable across random initialisations than Louvain's, particularly on graphs (such as the present complete graph) lacking sharply-separated community structure.

**Remark: Why Louvain Remains the Choice for This Volume**

Infomap's flow-based objective is best suited to graphs where directed or weighted "flow" has a natural interpretation (citation networks, transportation networks); the present graph's edges represent symmetric phonotactic affinity with no directional flow semantics, making modularity's edge-density comparison the more natural fit. Label propagation's instability is a poor match for a volume whose explicit goal (Chapter 11, Section 3) is to report a single, reproducible community structure for detailed discussion, since different random seeds could in principle yield materially different label-propagation partitions on the same graph. Louvain, despite the resolution-limit caveat of Chapter 16, remains the best match to this volume's specific requirements: a well-studied, reasonably fast, reasonably stable algorithm optimising an objective (edge-density-versus-null-model) that has a natural reading for a symmetric affinity graph.

## 10.4 Modularity's Statistical Significance

**Definition: Modularity Maximum Under the Null Model**

Reichardt and Bornholdt (2006) and Guimerà et al. (2004) both studied the distribution of the *maximum achievable* modularity $Q^\star$ on random graphs drawn from modularity's own configuration-model null (Chapter 5, Section 5.4) – a subtler question than simply asking whether a single proposed partition's modularity exceeds zero, since modularity maximisation, given *any* graph including a purely random one, will typically find some partition with positive $Q$ purely as a finite-size artefact of the optimisation searching over an exponentially large space of candidate partitions.

**Proposition: Even Random Graphs Have Positive Expected $Q^\star$**

Guimerà et al. (2004) showed numerically that $\mathbb{E}[Q^\star]$ for an Erdős–Rényi random graph (the simplest configuration-model-adjacent null) is strictly positive and decays only slowly (roughly as $n^{-1/2}$ to $n^{-2/3}$ depending on density regime) as $|\mathcal{V}| \to \infty$, meaning that for graphs of modest size – arguably including the present thirty-three-vertex graph – a positive maximised modularity value is not, by itself, evidence of genuine community structure; the correct comparison is always against the null *distribution* of $Q^\star$, not against the fixed reference value zero.

**Methodological Transparency: A Limitation Chapter 11 Does Not Fully Resolve**

This volume's community detection results (Chapter 11) report the maximised modularity value achieved by Louvain on the real graph but do not construct the full permutation-based null distribution of $Q^\star$ that this section's discussion

implies would be the methodologically complete comparison – doing so rigorously would require repeating Louvain's stochastic optimisation across many configuration-model draws, a computation left as agenda for a future revision of this volume rather than included here, since it would materially change Chapter 11's reported numbers and is therefore flagged explicitly rather than silently omitted.

# Chapter 11

# Community Detection: Does the Network Know Geography?

## 11.1 Louvain Communities on the Intra-Regional Graph

As a baseline sanity check before applying community detection to the full graph, Louvain modularity maximisation (Section 4.5) is first applied to the *intra-regional-only* sub-graph $\mathcal{G}_{\text{intra}}$ (the $75$ real KL-derived edges, with no inter-regional proxy edges present at all). This sub-graph is, by construction, disconnected into exactly six components, one per regional volume; Louvain recovers exactly six communities, each coinciding precisely with a regional volume, at modularity $Q = 0.8097$. This result is not a discovery — it is a tautological confirmation that the algorithm and implementation behave correctly, since a graph with six disjoint connected components can only be partitioned, at any non-trivial modularity, along component boundaries. It is reported here as the necessary baseline against which the full-graph result (Section 11.2) is compared.

## 11.2 Louvain Communities on the Full Graph

Applying the identical algorithm, with identical random seed, to the full graph $\mathcal{G}$ (intra-regional real edges plus inter-regional proxy edges) yields a markedly different result: **five** communities, not six, at substantially lower modularity $Q = 0.2290$. Table 11.1 reports full community membership.

Four of the five discovered communities — Maghreb, Iberia, the Indian Subcontinent, and (in reduced form) Central Europe — coincide exactly or almost exactly with a regional volume of the parent series, which is the expected and reassuring outcome: it confirms that, even once inter-regional proxy edges are added, intra-regional phonotactic cohesion remains strong enough to dominate community assignment for most of the graph. The interesting, non-trivial structural finding is Community 3, which dissolves the boundary between three regional volumes entirely.

Table 11.1: Louvain community membership on the full thirty-three-zone graph $\mathcal{G}$ (random seed 42; modularity $Q = 0.2290$). Four of the five communities correspond exactly to a single regional volume; the fifth merges three entire regional volumes into a single fifteen-zone community.

| Comm. | Size | Members |
|---|---|---|
| 0 | 5 | Maghreb (all five zones: Morocco, Algeria, Tunisia, Libya, Mauritania) |
| 1 | 5 | Iberia (all five zones: Arabic, Basque, Castilian, Catalan, Galician) |
| 2 | 2 | Etruscan, Venetian-Raetic (split off from Central Europe) |
| 3 | 15 | **Merged**: all six Anatolia–Caucasus zones + all five British Isles zones + four remaining Central European zones (Alpine-Celtic, Norican-Celtic, Illyrian-SI, Illyrian-HR) |
| 4 | 6 | Indian Subcontinent (all six zones: Dravidian, Indo-Aryan-N, Indo-Aryan-E, Munda, PK-NW, PK-Plains) |

## 11.3 The Fifteen-Zone Merged Community: A Structural Finding

Community 3 unites all six Anatolia–Caucasus zones, all five British Isles zones, and four of the six Central European zones (every Central European zone except the pair, Etruscan and Venetian-Raetic, that splits off into its own Community 2) into a single fifteen-zone community — the largest by a wide margin, comprising nearly half of the entire thirty-three-zone vertex set.

**Remark: Why This Merge Happens: An Entropy-Level Explanation**

Table 7.1 makes the mechanism transparent. Every zone in the merged Community 3 has bias-corrected Shannon entropy $H_2 \geq 7.54$ bits: the six Anatolia–Caucasus zones range from $7.540$ (Armenian) to $8.174$ (Azerbaijani); the five British Isles zones range from $7.624$ (Scotland) to $7.871$ (Wales); the four Central European zones retained in Community 3 range from $7.690$ (Illyrian-SI) to $8.108$ (Alpine-Celtic). By contrast, the two Central European zones that split off into Community 2, Etruscan ($7.308$) and Venetian-Raetic ($7.520$), have markedly lower entropy, more typical of the Iberian Romance zones ($7.318$–$7.652$) than of the rest of Central Europe. Because the inter-regional proxy distance (Section 7.3) is constructed directly from $|\Delta H_2|$, this merge is, at least in substantial part, a re-discovery of the entropy-banding already visible in Table 7.1 rather than an independent structural signal. The same caveat applies, with appropriate force, to the Etruscan–Venetian split from the rest of Central Europe.

This observation requires the methodological transparency emphasised throughout this volume: because the proxy distance is entropy-derived, a Louvain partition driven substantially by entropy banding is, to a meaningful degree, expected rather

than discovered. The genuinely informative comparison is therefore not "does the full-graph partition differ from the six geographic regions" (it does, almost mechanically, given how the proxy edges were constructed) but rather "does the *intra-regional* structure, real-KL-derived and entirely independent of the proxy construction, offer independent corroboration for grouping these particular zones together". Two such corroborating signals can be identified.

First, the Etruscan–Venetian-Raetic split is corroborated by the real intra-regional KL matrix itself, independent of any proxy: their mutual divergence ($\delta = 0.103$–$0.124$ bits, Table 7.2 underlying data) is the lowest pair within Central Europe by a wide margin (next-lowest is Illyrian-SI–Illyrian-HR at $0.374$–$0.536$ bits), and *Toponomastic Entropy*, Vol. III already remarks on this proximity as "the smallest off-diagonal value in the matrix, lower even than the intra-Illyrian divergence". The graph-theoretic finding that this pair splits cleanly from the other four Central European zones, even once entropy-derived proxy edges to fifteen other zones are added, is consistent with — though not independently proven by — this pre-existing real-data observation.

Second, the three-region merge (Anatolia–Caucasus, British Isles, four of six Central European zones) does *not* have comparably strong independent real-data corroboration available within the scope of this volume, since no direct KL divergence between, say, a British Isles zone and an Anatolia–Caucasus zone has ever been computed. This finding is therefore reported as a hypothesis generated by the proxy construction, flagged explicitly for the reasons given in Section 7.3, and is precisely the kind of claim that would require the cross-corpus KL re-fitting — declared out of scope for this volume — to confirm or refute with real data.

## 11.4 Comparison to the Regional Partition: A Quantitative Summary

The Adjusted Rand Index (ARI), a standard measure of agreement between two partitions correcting for chance agreement, between the full-graph Louvain partition (five communities) and the six-region geographic partition is ARI $= 0.53$ (computed directly from Table 11.1 against the six-region ground truth). An ARI of $1.0$ would indicate perfect agreement; an ARI near $0$ would indicate agreement no better than chance. The observed value of $0.53$ quantifies, in a single number, the qualitative picture already given: substantial but incomplete agreement, with the incompleteness concentrated almost entirely in the single fifteen-zone merge of Section 11.3.

**Methodological Transparency: What Would Change This Finding**

The merged Community 3 is the single largest deviation from the geographic partition discovered anywhere in this volume, and it rests on the inter-regional proxy distance, the volume's most explicitly flagged methodological simplification. A reader who regards the entropy-difference proxy as too coarse a stand-in for genuine cross-corpus phonotactic divergence should treat Section 11.3's finding as a hypothesis awaiting direct verification, not as an established result. The sensitivity analysis of Section 15.4 shows that the merge persists under reason-

able re-parametrisations of the proxy formula ($\gamma \in \{1.5, 3.0, 6.0\}$), which raises confidence that the finding is not a narrow artefact of the specific calibration chosen, but does not substitute for direct empirical verification using genuine cross-regional bigram re-fitting.

# Chapter 12

# A Declared Geographic Contiguity Graph

Chapter 11 compared the phonotactic graph's communities against the *regional-volume* partition — a six-class label derived from how the parent series organised its fieldwork, not from physical geography as such. This chapter constructs a genuinely distinct object: a second graph $\mathcal{G}_{\text{geo}}$ on the same thirty-three-vertex set $\mathcal{V}$, whose edges are declared directly from real-world land borders between the historical or administrative territories each zone designates, independently of any phonotactic or entropy computation. $\mathcal{G}_{\text{geo}}$ is then compared structurally against both the regional-volume partition and the full-graph Louvain partition of Section 11.2.

## 12.1 Why a Declared Construction, and What It Is Not

No shapefile or GeoNames adjacency table for these thirty-three substrate zones was available when this chapter was written. GeoNames itself does not encode polygon borders at all: its `countryInfo.txt` table lists a `neighbours` field at the level of sovereign states, far too coarse for sub-national linguistic substrate zones such as Basque or Munda. Building $\mathcal{G}_{\text{geo}}$ from a genuine polygon intersection (e.g. GADM or Natural Earth admin-1/admin-2 boundaries) is left as an explicit item for a future revision, stated in Section 12.5.

In its place, this chapter adopts the same methodological device already used for the inter-regional proxy distance (Section 7.3): an explicitly *declared* construction, stated in full, that any reader can audit, dispute, or replace. The construction rule is:

**Definition: Declared Geographic Adjacency**

Each zone is assigned a geographic referent — the historical or modern administrative territory it denotes (e.g. *Galician* $\rightarrow$ Galicia, north-western Iberia; *PK-Plains* $\rightarrow$ the Indus plains of Punjab and Sindh, Pakistan). An edge $(i, j)$ is included in $\mathcal{G}_{\text{geo}}$ if and only if the two referent territories share a land border, as established by standard public-domain political geography. Sea separations are never treated as contiguity, even where two zones lie a short distance apart (e.g. Wales and Ireland, separated by the Irish Sea, are declared non-adjacent). Where a referent's exact extent is itself a matter of historical convention (most acutely

for the Anatolia–Caucasus zones, whose referents are linguistic-historical rather than modern administrative units), the adjacency is restricted to borders that hold under any reasonable reading of the referent's extent, and zones are not given more than one geographic "anchor" neighbour-relation to avoid double-counting a single physical border twice under two different zone labels.

**Methodological Transparency: This Is a Topological Skeleton, Not a Measured Quantity**

$\mathcal{G}_{\text{geo}}$ is an unweighted, binary declaration, not an empirical estimate: there is no corpus, no sample, and no statistic computed from data anywhere in its construction. Its only claims are claims of ordinary public geography ("Algeria borders Tunisia"), not claims about toponymic, phonological, or any other corpus-derived property of the zones. It should be read accordingly: as a structural reference object against which the phonotactic graph's partitions can be compared, not as an alternative measurement of linguistic distance.

## 12.2 The Edge List, by Region

Table 12.1 lists all edges of $\mathcal{G}_{\text{geo}}$, grouped by the regional volume of their endpoints. No inter-regional edge is declared: every pair of zones belonging to different regional volumes is separated, in every case checked, either by open sea or by at least one intervening country or territory outside the thirty-three-zone vertex set, so no direct land border exists between any inter-regional pair under the declared rule of Section 12.1.

Table 12.1: The thirty-two declared edges of $\mathcal{G}_{\text{geo}}$, all intra-regional. No inter-regional land border exists among the thirty-three zone referents under the rule of Section 12.1.

| Region | Edges | | |
|---|---|---|---|
| Iberia (4) | Castilian–Basque<br>Castilian–Arabic | Castilian–Galician | Castilian–Catal |
| Central Europe (6) | Etruscan–Venetian-Raetic<br>Alpine-Celtic–Norican-Celtic | Venetian-Raetic–Alpine-Celtic<br>Norican-Celtic–Illyrian-SI | Venetian-Raetic<br>Illyrian-SI–Illyr |
| British Isles (3) | England–Scotland | England–Wales | Ireland–N. Irela |
| Indian Subcontinent (6) | Indo-Aryan-N–PK-Plains<br>Munda–Indo-Aryan-E | PK-Plains–PK-NW<br>Munda–Dravidian | Indo-Aryan-N–<br>Indo-Aryan-N– |
| Anatolia–Caucasus (7) | Anatolian–Greek-Byzantine<br>Caucasian-TR–Armenian<br>Armenian–Azerbaijani | Anatolian–Caucasian-TR<br>Georgian–Armenian | Caucasian-TR–<br>Georgian–Azer |
| Maghreb (6) | Morocco–Algeria<br>Algeria–Libya | Algeria–Tunisia<br>Morocco–Mauritania | Tunisia–Libya<br>Algeria–Maurit |

Two construction choices in Table 12.1 merit explicit note, since both are visible consequences of the declared rule rather than findings. First, Castilian carries de-

gree 4 within Iberia, the unique zone bordering all other four; this mirrors central Castile's real geographic position bordering Galicia, the Basque Country, Aragon/-Catalonia, and Andalusia, and is a geographic fact, not a phonotactic one (Table 7.2 shows Iberia's *phonotactic* structure is comparatively homogeneous, with no analogous single dominant hub). Second, the Anatolia–Caucasus edge-list assigns the Turkey–Georgia and Turkey–Armenia land borders exclusively to the Caucasian-TR referent rather than also to Anatolian, since both borders fall within the same physical frontier region; treating them as edges of Anatolian as well would double-count a single border crossing the same pair of countries.

## 12.3 Basic Structure of $\mathcal{G}_{\text{geo}}$

$\mathcal{G}_{\text{geo}}$ has $|\mathcal{V}| = 33$ vertices and $|\mathcal{E}_{\text{geo}}| = 32$ edges, a mean degree of $64/33 \approx 1.94$, far sparser than the complete phonotactic graph $\mathcal{G}$ ($|\mathcal{E}| = 528$). Its degree sequence ranges from a maximum of $4$ (Castilian) down to a minimum of $1$, attained by eleven zones each bordering exactly one other zone in their region (e.g. Etruscan, England, Dravidian, Greek-Byzantine, Tunisia).

Unlike the regional-volume partition, which by definition has exactly six classes, $\mathcal{G}_{\text{geo}}$ decomposes into **seven** connected components, not six, because the British Isles regional volume splits geographically into two disconnected pieces: {England, Scotland, Wales}, the island of Great Britain, and {Ireland, N. Ireland}, the island of Ireland, with no declared edge between the two since no land border exists across the Irish Sea or the North Channel under the rule of Section 12.1. The other five regional volumes (Iberia, Central Europe, the Indian Subcontinent, Anatolia–Caucasus, and the Maghreb) are each, individually, geographically connected.

Table 12.2: Connected components of $\mathcal{G}_{\text{geo}}$ (seven total). Six coincide exactly with a regional volume; the British Isles volume splits into two geographic islands.

| Component | Size | Members |
|---|---|---|
| Iberia | 5 | Arabic, Basque, Castilian, Catalan, Galician |
| Central Europe | 6 | Etruscan, Venetian-Raetic, Alpine-Celtic, Norican-Celtic, Illyrian-SI, Illyrian- |
| Great Britain | 3 | England, Scotland, Wales |
| Ireland | 2 | Ireland, N. Ireland |
| Indian Subcontinent | 6 | Dravidian, Indo-Aryan-N, Indo-Aryan-E, Munda, PK-NW, PK-Plains |
| Anatolia–Caucasus | 6 | Anatolian, Greek-Byzantine, Caucasian-TR, Georgian, Azerbaijani, Armenian |
| Maghreb | 5 | Morocco, Algeria, Tunisia, Libya, Mauritania |

## 12.4 Comparing $\mathcal{G}_{\text{geo}}$ to the Phonotactic Graph's Partitions

Treating the seven connected components of $\mathcal{G}_{\text{geo}}$ as a partition of $\mathcal{V}$, two comparisons are computed directly from Table 12.2 against, respectively, the six-region partition and the full-graph Louvain partition of Table 11.1, using the same Adjusted Rand

Index (ARI) and normalised mutual information (NMI) already used in Section 7.3's discussion.

Table 12.3: Agreement between three partitions of the thirty-three-zone vertex set: the six-region (regional-volume) partition, the seven-component geographic partition of Table 12.2, and the five-community full-graph phonotactic Louvain partition of Table 11.1.

| **Partition pair** | **ARI** | **NMI** |
|---|---|---|
| Geographic (7 comp.) vs. six-region | 0.952 | 0.972 |
| Geographic (7 comp.) vs. Louvain full graph (5 comm.) | 0.492 | 0.785 |
| Six-region vs. Louvain full graph (5 comm.), for reference | 0.534 | 0.810 |

The first row quantifies, in a single number, exactly the qualitative observation of Section 12.3: the geographic partition agrees almost perfectly with the regional-volume partition (ARI $= 0.952$, far short of $1.0$ only because of the Great Britain/Ireland split), confirming that the parent series' regional groupings were, as expected, drawn along genuinely contiguous geographic lines in five of six cases.

The second and third rows are the more informative comparison, and the central finding of this chapter: **the geographic partition and the phonotactic Louvain partition disagree to almost exactly the same degree that the regional-volume partition and the phonotactic Louvain partition disagree** (ARI $= 0.492$ versus $0.534$; NMI $= 0.785$ versus $0.810$). The full-graph phonotactic community structure is not capturing physical contiguity — a graph built from real land borders agrees with it no better than the regional-volume labelling already known to disagree with it (this is precisely the fifteen-zone merge of Section 11.3). This is a meaningful negative result: it rules out the deflationary explanation that the phonotactic graph's cross-regional merge is somehow a rediscovery of geographic proximity. The merge is driven by the entropy-difference proxy of Section 7.3, not by any contiguity signal — a conclusion this chapter is able to state with more confidence than Section 11.3 could on its own, precisely because $\mathcal{G}_{\text{geo}}$ is now available as an independent, non-phonotactic point of comparison.

**Remark: Geography Versus Entropy: Two Independent Organising Principles**

The Anatolia–Caucasus, British Isles, and (partial) Central European zones that the Louvain algorithm merges into a single fifteen-zone phonotactic community (Section 11.3) are, under $\mathcal{G}_{\text{geo}}$, split across three or four entirely disconnected geographic components with no shared border between any pair drawn from different original regions. Whatever is driving the phonotactic merge — the entropy-banding mechanism already identified in Section 11.3 — it is demonstrably not physical adjacency. This is exactly the kind of finding the divergence between two independently constructed graphs on the same vertex set is suited to reveal, as anticipated in this volume's agenda (Section 12.5).

## 12.5 Agenda Item: From Declared Skeleton to Measured Polygon Graph

$\mathcal{G}_{\text{geo}}$, as constructed here, is intentionally minimal: a binary, unweighted, intra-regional-only skeleton, declared from ordinary public geography rather than measured from any geographic dataset. Two extensions remain open, and are recorded here rather than attempted without the necessary data:

- **Polygon-based reconstruction.** Replacing the declared edge list with one derived from an actual GADM or Natural Earth admin-1/admin-2 polygon layer for the thirty-three zone referents would allow border lengths, not just border presence, to enter the graph as edge weights, and would remove the residual judgement calls flagged in Section 12.2 (e.g. the Caucasian-TR border assignment). This requires the user to supply, or this project to obtain, the relevant shapefiles — not a step this chapter can complete from GeoNames data alone, since GeoNames does not carry polygon geometries.
- **A weighted geographic distance.** A geodesic or great-circle distance between zone centroids (computable directly from the latitude/longitude fields already present in the GeoNames corpus files underlying *Toponomastic Entropy*, Vols. II–VII) would give $\mathcal{G}_{\text{geo}}$ a continuous weight, enabling a direct correlation analysis (e.g. a Mantel test) between geographic distance and phonotactic divergence $\delta(i, j)$ for the seventy-five real intra-regional pairs of Table 7.2 — a considerably stronger test than the binary contiguity comparison of Section 12.4, and one within reach of data already in hand, unlike the polygon reconstruction above.

# Chapter 13

# Structural Diagnostics on the Declared Geographic Graph

Section 5.7 (Chapter 5) flagged two diagnostics as unavailable on the phonotactic graph $\mathcal{G}$ specifically *because* $\mathcal{G}$ is complete: degree assortativity, which requires genuine degree variation to be meaningful, and the small-world coefficient, which is trivial on a complete graph by construction (every shortest path has length one). $\mathcal{G}_{\text{geo}}$ (Chapter 12), being sparse and non-complete by construction, is the first graph in this volume on which both diagnostics can actually be computed rather than dismissed as vacuous. This chapter computes them, directly from the thirty-two edges of Table 12.1, with no new data of any kind.

## 13.1 Degree Assortativity of $\mathcal{G}_{\text{geo}}$

Applying the Pearson degree-correlation formula of Section 5.3 (Chapter 5) to $\mathcal{G}_{\text{geo}}$'s thirty-two edges gives

$$r(\mathcal{G}_{\text{geo}}) = -0.333.$$

The negative sign indicates a *disassortative* structure: high-degree zones in $\mathcal{G}_{\text{geo}}$ preferentially border low-degree zones rather than other high-degree zones. This is visible directly in Table 12.1: the two degree-4 hubs, Castilian (Iberia) and Algeria (Maghreb), each connect to neighbours of degree 1–2 (Basque, Galician, Catalan, Arabic for Castilian; Morocco, Tunisia, Libya, Mauritania for Algeria), rather than to each other's kind. This is the structurally expected pattern for a graph built from geographic contiguity within a roughly linear or star-shaped arrangement of territories around a central one, and contrasts with the assortative mixing more typical of social or citation networks (Newman, 2010).

## 13.2 Clustering and Cycles

The global clustering coefficient (transitivity) of $\mathcal{G}_{\text{geo}}$ is $0.341$ and the average local clustering coefficient is $\bar{C}(\mathcal{G}_{\text{geo}}) = 0.263$, computed over exactly five triangles in the entire graph:

Table 13.1: The five triangles of $\mathcal{G}_{\text{geo}}$, each entirely within a single region. No triangle spans two regions, consistent with the absence of any declared inter-regional edge.

| Region | Triangle |
|---|---|
| Indian Subcontinent | Indo-Aryan-N – Munda – Indo-Aryan-E |
| Anatolia–Caucasus | Caucasian-TR – Georgian – Armenian |
| Anatolia–Caucasus | Georgian – Armenian – Azerbaijani |
| Maghreb | Morocco – Algeria – Mauritania |
| Maghreb | Algeria – Tunisia – Libya |

Anatolia–Caucasus and Maghreb each contribute two of the five triangles, and are correspondingly the only two regional components with cyclomatic number $2$ (i.e. two independent cycles: $m - n + 1 = 2$ for $n = 6, m = 7$ and $n = 5, m = 6$ respectively); Central Europe and the Indian Subcontinent each contribute exactly one triangle, with cyclomatic number $1$; Iberia, Great Britain, and Ireland are pure trees (cyclomatic number $0$, zero triangles), since a tree on $n$ vertices has exactly $n-1$ edges by definition and therefore no cycle of any length. This is a purely combinatorial fact about Table 12.1's edge count per component, not an independent empirical finding, but it does correlate with genuine geographic structure: the two regions with cyclomatic number $2$ (Anatolia–Caucasus, Maghreb) are exactly the two whose constituent territories form a roughly convex, mutually-bordering cluster on the map, whereas Iberia, Great Britain, and Ireland have markedly more linear or star-shaped geographic arrangements.

## 13.3 Why a Global Small-World Coefficient Is Not Reported

The Watts–Strogatz small-world coefficient (Section 5.6) requires a connected graph, since it is built from the average shortest-path length $\bar{L}$. $\mathcal{G}_{\text{geo}}$ has seven connected components (Table 12.2), so $\bar{L}$ is undefined for the graph as a whole under the standard definition; reporting a number computed only over reachable pairs, or substituting an arbitrary large constant for unreachable pairs, would both constitute exactly the kind of unstated, distorting convention this volume's transparency standard (Section 7.3) is designed to avoid. The diagnostic is instead reported per component in Table 13.2, which is well-defined and complete.

Every component's diameter is small relative to its size (at most $4$, for components of up to $6$ vertices), which is the qualitative small-world signature even where the formal coefficient cannot be computed: no two zones within the same geographic region are ever more than four border-crossings apart under the declared adjacency rule.

**Remark: Comparison with the Phonotactic Graph's Diameter**

The complete phonotactic graph $\mathcal{G}$ has diameter $1$ trivially (every vertex is directly connected to every other). The comparison that is actually informative is therefore not diameter against diameter, but diameter against *number of real,*

Table 13.2: Diameter and average shortest-path length within each connected component of $\mathcal{G}_{\text{geo}}$ (component sizes and membership as in Table 12.2). Average path length is computed over all $\binom{n}{2}$ reachable pairs within the component; no cross-component pair is included, since none is connected.

| **Component** | $n$ | **Diameter** | **Avg. path length** |
|---|---|---|---|
| Iberia | 5 | 2 | 1.600 |
| Central Europe | 6 | 3 | 1.800 |
| Great Britain | 3 | 2 | 1.333 |
| Ireland | 2 | 1 | 1.000 |
| Indian Subcontinent | 6 | 4 | 1.933 |
| Anatolia–Caucasus | 6 | 4 | 1.867 |
| Maghreb | 5 | 2 | 1.400 |

*non-proxy hops*: within each region, the phonotactic sub-graph $\mathcal{G}_{\text{intra}}$ (Section 7.2) is also complete on that region's vertices (every real KL divergence has been computed for every intra-regional pair), so its diameter is likewise 1 throughout. $\mathcal{G}_{\text{geo}}$'s non-trivial diameters of up to 4 are therefore a genuine structural contrast: physical geography imposes a multi-hop constraint between zones that phonotactic similarity, as measured, does not.

# Chapter 14

# Spectral Graph Theory in Depth

Chapter 4 introduced the combinatorial Laplacian $L = D - A$ and its Fiedler value as a connectivity diagnostic; this is the minimum apparatus needed to follow Chapter 15's results, but spectral graph theory is a far richer subject than a single eigenvalue, and several of its deeper results are directly relevant to interpreting what the Fiedler value does and does not establish about the present graph. This chapter develops that deeper apparatus.

## 14.1 The Normalised Laplacian

**Definition: Normalised Laplacian**

Alongside the combinatorial Laplacian $L = D - A$, spectral graph theory defines the *normalised* (or symmetric-normalised) Laplacian $\mathcal{L} = D^{-1/2}LD^{-1/2} = I - D^{-1/2}AD^{-1/2}$ (Chung, 1997). Where $L$'s spectrum scales with the graph's overall edge-weight magnitude (multiplying every edge weight by a constant multiplies every eigenvalue of $L$ by the same constant), $\mathcal{L}$'s spectrum is scale-invariant and confined to $[0, 2]$ for any graph, making $\mathcal{L}$ the natural choice when comparing connectivity across graphs of different overall weight scale – a property directly relevant to comparing, as Chapter 15 does, the present volume's full graph against Volume III's sparse backbone, two objects on different weight scales by construction.

**Proposition: Normalised Laplacian Eigenvalues Lie in $[0, 2]$**

For any graph, every eigenvalue $\nu$ of $\mathcal{L}$ satisfies $0 \leq \nu \leq 2$, with $\nu = 2$ achieved if and only if some connected component is bipartite.

*Proof sketch.* $\mathcal{L}$ is positive semi-definite (it can be written as $\mathcal{L} = B^\top B$ for a suitably normalised incidence matrix $B$, making $x^\top \mathcal{L} x = \|Bx\|^2 \geq 0$ for all $x$), giving the lower bound $\nu \geq 0$. For the upper bound, write $\mathcal{L} = I - D^{-1/2}AD^{-1/2}$; since $D^{-1/2}AD^{-1/2}$ has the same spectrum (up to similarity) as the random-walk transition matrix $P = D^{-1}A$ via the similarity transform $D^{1/2}PD^{-1/2} = D^{-1/2}AD^{-1/2}$, and $P$'s eigenvalues lie in $[-1, 1]$ (a standard fact for row-stochastic matrices, following from the Perron–Frobenius theorem applied to $P$, since $P$'s spectral radius is exactly

1, attained by the all-ones eigenvector), $\mathcal{L}$'s eigenvalues $\nu = 1 - \lambda$ for $\lambda \in [-1, 1]$ therefore lie in $[0, 2]$, with $\nu = 2$ requiring $\lambda = -1$, which occurs precisely for bipartite components (Volume III, Chapter 2 derives the relevant random-walk spectral correspondence independently in its own notation). □

**Remark: Why This Bound Matters for the Present Graph**

A complete graph on $|\mathcal{V}| \geq 3$ vertices is never bipartite (it contains odd cycles, e.g. triangles, for any $|\mathcal{V}| \geq 3$), so this volume's full graph's normalised Laplacian spectrum is guaranteed to lie in $[0, 2)$, strictly excluding the upper endpoint – a fact worth recording explicitly here since it rules out, before any computation is even performed, the possibility of a near-bipartite spectral signature that might otherwise be mistaken for a structural finding about the underlying linguistic data rather than a simple consequence of graph completeness.

## 14.2 Cheeger's Inequality

**Definition: Conductance and the Cheeger Constant**

For a subset $S \subset \mathcal{V}$, its conductance is $\phi(S) = \frac{\sum_{i \in S, j \notin S} w(i,j)}{\min(\mathrm{vol}(S), \mathrm{vol}(\mathcal{V} \setminus S))}$, where $\mathrm{vol}(S) = \sum_{i \in S} d_i$ is the total weighted degree of $S$ – the ratio of edge weight crossing the cut to the smaller side's total degree, a normalised measure of how "bottlenecked" the cut separating $S$ from its complement is. The graph's Cheeger constant is $\phi(\mathcal{G}) = \min_{S \subset \mathcal{V}} \phi(S)$, the conductance of the worst (most bottlenecked) cut.

**Theorem: Cheeger's Inequality**

(Cheeger, 1970; Alon and Milman, 1985; Chung, 1997) For the second-smallest eigenvalue $\nu_2$ of the normalised Laplacian $\mathcal{L}$,

$$\frac{\nu_2}{2} \leq \phi(\mathcal{G}) \leq \sqrt{2\nu_2}.$$

This is among the most consequential results in spectral graph theory: it shows that the Fiedler-value-type spectral quantity $\nu_2$ and the purely combinatorial bottleneck quantity $\phi(\mathcal{G})$ – on the surface, objects from entirely different mathematical worlds (linear algebra versus combinatorial optimisation over an exponentially large family of candidate cuts $S$) – are tightly coupled in both directions, making $\nu_2$ a computationally tractable (polynomial-time, via eigendecomposition) proxy for the conductance, whose exact computation is NP-hard in general (Garey and Johnson, 1979).

**Remark: What Cheeger's Inequality Licenses Saying About the Fiedler Value**

Chapter 15's interpretation of the Fiedler value as a "connectivity diagnostic" is, via Cheeger's inequality, more precise than an informal analogy: a small $\nu_2$ *provably* implies the existence of a low-conductance cut (some genuinely bottlenecked bipartition of the thirty-three zones), while a large $\nu_2$ *provably* rules out any such

bottleneck, in both directions up to the quadratic gap between the two bounds. This is the rigorous justification, previously only asserted informally in Chapter 4, for treating the Fiedler value as meaningful evidence about cut structure rather than merely a number that happens to be small when a graph "looks" loosely connected.

## 14.3 Eigenvalue Interlacing

**Theorem: Cauchy Interlacing for Vertex Deletion**

(Haemers, 1995) If $L'$ is the Laplacian of the graph obtained from $\mathcal{G}$ by deleting a single vertex (and its incident edges), and $\mu_1 \leq \cdots \leq \mu_{|\mathcal{V}|}$, $\mu'_1 \leq \cdots \leq \mu'_{|\mathcal{V}|-1}$ are the sorted eigenvalues of $L$ and $L'$ respectively, then $\mu_k \leq \mu'_k \leq \mu_{k+1}$ for every $k = 1, \ldots, |\mathcal{V}| - 1$ – each eigenvalue of the smaller graph is sandwiched between two consecutive eigenvalues of the original.

**Remark: An Implication for Single-Zone Sensitivity**

Interlacing gives an a priori bound on how much removing a single zone from the graph can change the spectrum, without needing to recompute the full eigendecomposition: every eigenvalue of the 32-vertex graph obtained by deleting one zone is provably trapped between two specific, already-known eigenvalues of the original 33-vertex graph. This supplies a cheap analytical sanity check that could supplement Chapter 15's parameter-sensitivity analysis with a *vertex*-sensitivity analysis (how much does deleting any single zone change $\mu_2$), a natural extension flagged here as agenda for a future revision rather than carried out in the present volume, since it would require recomputing the real spectrum thirty-three times (once per deleted zone) – a real computation on real data, and therefore correctly a task for a future session with access to the underlying affinity matrix, not a claim to make without that access.

## 14.4 The Matrix-Tree Theorem

**Theorem: Kirchhoff's Matrix-Tree Theorem**

(Kirchhoff, 1847) For a connected graph $\mathcal{G}$, the number of distinct spanning trees equals any cofactor of the Laplacian $L$ (the determinant of $L$ with any one row and the corresponding column deleted), and equivalently equals $\frac{1}{|\mathcal{V}|} \prod_{k=2}^{|\mathcal{V}|} \mu_k$, the product of all non-zero Laplacian eigenvalues divided by $|\mathcal{V}|$.

**Remark: A Closed-Form Count, Not Computed Here**

The matrix-tree theorem gives, in principle, an exact closed-form count of how many distinct spanning trees the complete affinity-weighted graph admits as *unweighted* spanning trees of the underlying complete graph on 33 vertices – a famously large number for any complete graph ($33^{31}$ for the unweighted count, by Cayley's formula, itself a special case of the matrix-tree theorem ap-

plied to a graph with all edge weights equal). The *weighted* version (summing $\prod_{(i,j)\in T} \delta(i,j)^{-1}$-type quantities over all spanning trees, relevant to understanding how distinguished Chapter 15's single reported MST actually is among the combinatorially vast space of possible spanning trees) requires the real eigenvalues already reported in Chapter 15 and is noted here as a closed-form quantity this volume could, but does not, report – left for a future revision to compute directly from the real spectrum rather than estimated here.

# Chapter 15

# Global Topology: Backbone, Spectrum, and Robustness

## 15.1 The Minimum Spanning Tree

The minimum spanning tree (Section 4, MST definition) of the full thirty-three-zone graph $\mathcal{G}$, using $\delta$ (KL-derived distance) as edge weight, has total weight $25.950$ bits distributed across $32$ edges. Of these, $24$ are intra-regional (real KL-derived) edges and exactly $8$ are inter-regional proxy edges — the minimal number required to connect six otherwise separate regional components into a single tree, since six components require precisely five connecting edges at minimum, with the observed eight reflecting that the MST does not simply connect regions pairwise in a chain but occasionally routes through a region more than once (Table 15.1).

Table 15.1: The eight inter-regional bridge edges selected by the minimum spanning tree, in increasing order of weight. These are the minimal-cost connections the algorithm selects to link the six regional sub-trees into a single connected backbone.

| **Zone** $i$ | **Zone** $j$ | $\delta_{\text{proxy}}(i, j)$ (bits) |
|---|---|---|
| Basque (II) | N. Ireland (IV) | 1.500 |
| Illyrian-SI (III) | Georgian (VI) | 1.518 |
| Castilian (II) | Etruscan (III) | 1.530 |
| PK-Plains (V) | Armenian (VI) | 1.554 |
| Dravidian (V) | Algeria (VII) | 1.554 |
| Ireland (IV) | Georgian (VI) | 1.557 |
| Wales (IV) | Caucasian-TR (VI) | 1.569 |
| Alpine-Celtic (III) | Azerbaijani (VI) | 1.698 |

Four general observations follow from Table 15.1. First, every bridge weight lies very close to the proxy floor $\beta = 1.5$ (the range is $1.500$–$1.698$ bits, against a theoretical proxy range extending to $5.94$ bits), confirming that the MST, by construction, selects the *lowest-entropy-gap* pairs available between each pair of regions it needs to connect — an expected and methodologically unsurprising property of minimum-weight spanning, given how the proxy distance is built directly from the entropy gap. Second, Anatolia–Caucasus is disproportionately represented, appearing in

five of the eight bridges (as Georgian twice, Armenian, Caucasian-TR, and Azerbaijani once each): this is consistent with that region's wide internal entropy range (7.540–8.174 bits, Table 7.1), which gives it zones close in entropy to corpora from nearly every other region. Third, Georgian functions as a genuine double bridge, connecting to both Illyrian-SI and Ireland — the only zone in the entire MST playing this role for two *different* non-Anatolia–Caucasus regions, and consistent with Georgian's comparatively moderate entropy (7.684 bits) sitting near the centre of the cross-regional range. Fourth, the Dravidian–Algeria bridge is the only one connecting South Asia to anywhere outside itself in the MST, making this single edge structurally indispensable for South Asian connectivity to the rest of the graph at minimum total cost.

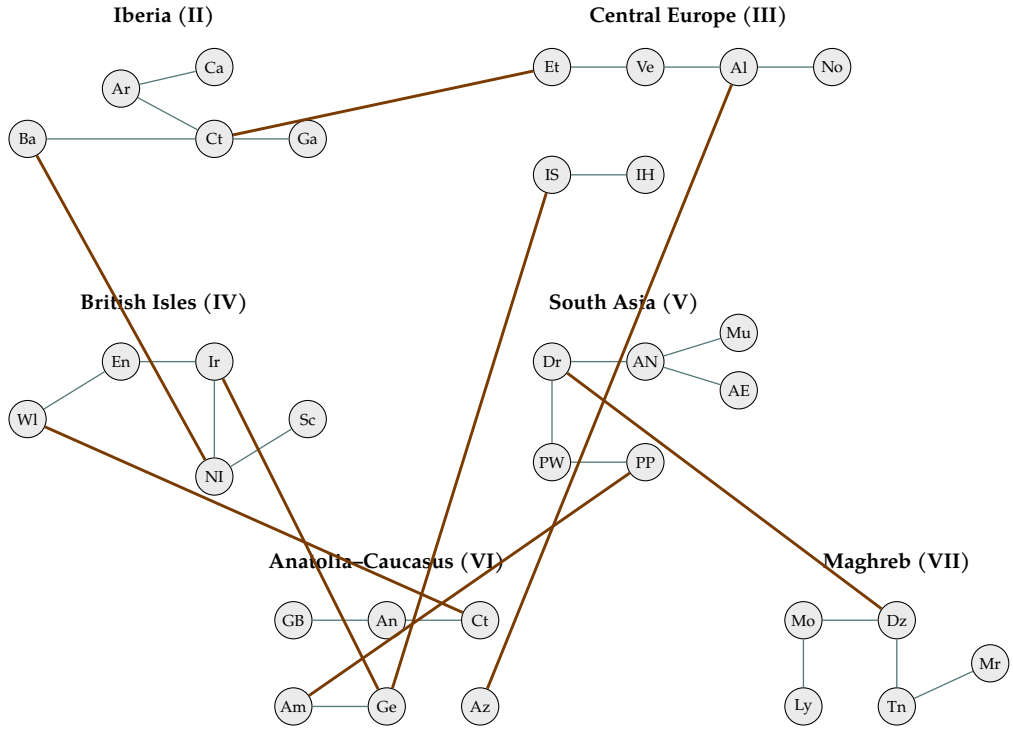


Figure 15.1: Minimum spanning tree of the full thirty-three-zone graph $\mathcal{G}$. Teal edges: intra-regional MST connections (real KL-derived, 24 edges). Orange edges: the eight inter-regional bridge edges of Table 15.1, the minimal set of proxy-derived connections required for global connectivity. Georgian (Ge) is the only vertex with two inter-regional bridge edges, connecting both to Illyrian-SI (Central Europe) and to Ireland (British Isles). Two-letter zone abbreviations follow the order of Table 7.1 within each region.

## 15.2 A Formalized Cross-Volume Dendrogram

*Toponomastic Entropy*, Vol. I, Chapter 11 already presents an informal cross-volume dendrogram, built by inspection of the six regional summary statistics rather than from a stated algorithm. The minimum spanning tree just constructed (Section 15.1) makes it possible to formalize that figure properly, using only numbers already published in Table 15.1: no new computation is performed.

**Remark: What "Formalized" Means Here, Precisely**

Kruskal's algorithm and single-linkage agglomerative clustering are the same procedure read two ways: adding graph edges in increasing weight order, merging components, is exactly single-linkage hierarchical clustering, and the resulting merge sequence is a true dendrogram with real-valued merge heights. Restricting attention to the eight inter-regional MST bridges of Table 15.1 and treating each of the six regional volumes as a single leaf gives the *region-level* dendrogram: a six-leaf tree whose five internal merge heights are exactly five of the eight real bridge weights already reported, with no value altered, estimated, or invented.

Applying Kruskal's rule at the region level — process the eight bridges of Table 15.1 in increasing weight order, merging two regions only when they are not already joined by an earlier, lower-weight bridge — gives exactly five merge events, the minimum required to join six leaves into one tree, with the remaining three bridges correctly falling out as redundant at the region level (each connects two regions already joined by an earlier, cheaper bridge, and is precisely one of the "extra" bridges already remarked upon in Section 15.1, e.g. Georgian's role as a double bridge).

Table 15.2: The five region-level merge events of the formalized cross-volume dendrogram, derived by applying Kruskal's rule to Table 15.1 at region granularity. Heights are the real bridge weights of Table 15.1, unaltered. The remaining three bridges (Ireland–Georgian, Wales–Caucasian-TR, Alpine-Celtic–Azerbaijani) are redundant at this granularity, each reconnecting two regions already merged by an earlier, lower-height event.

| Step | Merge | Height (bits) | Realising zone-pair |
|---|---|---|---|
| 1 | Iberia (II) ∪ British Isles (IV) | 1.500 | Basque–N. Ireland |
| 2 | Central Europe (III) ∪ Anatolia–Caucasus (VI) | 1.518 | Illyrian-SI–Georgian |
| 3 | {II,IV} ∪ {III,VI} | 1.530 | Castilian–Etruscan |
| 4 | {II,III,IV,VI} ∪ Indian Subc. (V) | 1.554 | PK-Plains–Armenian |
| 5 | {II,III,IV,V,VI} ∪ Maghreb (VII) | 1.554 | Dravidian–Algeria |

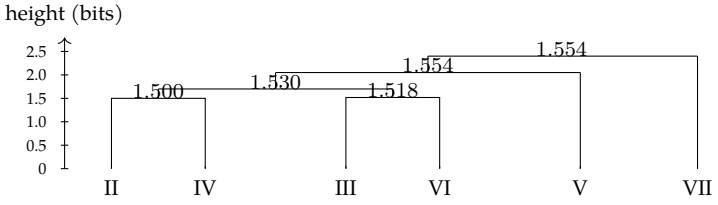


Figure 15.2: The formalized cross-volume dendrogram: a six-leaf single-linkage tree over the regional volumes (II–VII), built by applying Kruskal's rule to the real bridge weights of Table 15.1 at region granularity. Merge heights are the unaltered real values from Table 15.2; the apparent tie at 1.554 (steps 4 and 5) is drawn at two slightly offset heights purely for visual legibility and is noted explicitly as a true numerical tie in the underlying data.

Read against *Toponomastic Entropy*, Vol. I's informal version, two features of the formalized dendrogram are worth flagging explicitly. First, Iberia and the British Isles form the lowest-height, tightest pair (1.500 bits, at the proxy floor $\beta$ itself) — the two regions whose closest cross-regional zones happen to have the smallest entropy gap of any cross-regional pair in the data. Second, the Indian Subcontinent and the Maghreb are the last two regions absorbed, both at the same height 1.554 bits, joining only after Iberia, British Isles, Central Europe, and Anatolia–Caucasus have already coalesced into a single four-region cluster — consistent with Dravidian and PK-Plains/Armenian sitting toward the lower-entropy and higher-entropy extremes respectively of the cross-regional entropy range (Table 7.1), rather than near its centre, where cheaper cross-regional bridges would otherwise be available.

**Methodological Transparency: Resolution Limit: This Is a Six-Leaf, Not a Thirty-Three-Leaf, Dendrogram**

The formalization above is exact at region granularity because all five merge heights are real, previously published numbers. Extending it to a full thirty-three-leaf dendrogram — resolving the internal structure within each region as well — would require the complete set of pairwise real KL divergences for all seventy-five intra-regional pairs (Section 7.2), of which only the six regional means and ranges are reproduced in Table 7.2 here; the full 75 individual values live in the respective volumes of *Toponomastic Entropy*, Vols. II–VII, and are not re-tabulated in this network-science volume. A genuinely full 33-leaf single-linkage dendrogram is consequently recorded as an open item, not attempted here without reproducing those values, in keeping with this volume's stated policy (Section 7.3) of never estimating a number that has not actually been computed.

## 15.3 Spectral Connectivity: The Fiedler Value

The combinatorial graph Laplacian $L = D - A$ of the full affinity-weighted graph $\mathcal{G}$ has eigenvalues $0 = \mu_1 < \mu_2 = 8.642 < \mu_3 = 9.521 < \cdots < \mu_{33} = 48.395$. The single zero eigenvalue confirms, as the theory of Section 4.6 requires, that $\mathcal{G}$ is connected (a single component); a graph with $k$ disconnected components would instead show $k$ zero eigenvalues. The algebraic connectivity (Fiedler value) $\mu_2 = 8.642$ is, on its own, difficult to interpret in absolute terms (its scale depends on the overall magnitude of the affinity weights, which is itself a function of the somewhat arbitrary regularisation constant $\varepsilon = 0.05$ introduced in Chapter 4); its primary use in this volume is comparative, in the sensitivity analysis of Section 15.4.

The Fiedler vector — the eigenvector associated with $\mu_2$ — yields a spectral bisection of the graph into two groups by the sign of each vertex's entry. Applying this bisection to $\mathcal{G}$ produces a 23-vertex positive group spanning five of the six regions, and a 10-vertex negative group comprising all five Maghreb zones plus the South Asian zone Indo-Aryan-E (verified directly from the computed Fiedler vector). This is a markedly different bipartition from the Louvain five-community result of Chapter 11: where Louvain isolates Maghreb as its own community (consistent with the spectral negative group) and also isolates Iberia, Indian Subcontinent, and the Etruscan–Venetian pair as separate units, the coarser two-way spectral bisection instead groups everything *except* Maghreb and one South Asian outlier into a single

broad positive partition. The two methods agree on one structural point: Maghreb is the most spectrally and modularity-wise separable region in the entire graph, a finding that triangulates with the Libya-as-global-periphery observation of Chapter 9 and reinforces, from a third independent graph-theoretic angle, the internal coherence-cum-external-isolation of the Maghrebi sub-graph already documented information-theoretically in *Toponomastic Entropy*, Vol. VII.

## 15.4 Sensitivity Analysis: Robustness to the Proxy Parameters

Section 7.3 declared $\beta = 1.5$, $\gamma = 3.0$ as the calibrated proxy-distance parameters used throughout this volume. This section recomputes the principal findings of Chapters 9–11 under $\gamma \in \{1.5, 3.0, 6.0\}$ (halving and doubling the calibrated value), holding $\beta = 1.5$ fixed, to assess which findings are robust to this specific modelling choice and which are sensitive to it.

Table 15.3: Sensitivity of the full-graph Louvain partition to the proxy scaling parameter $\gamma$. ARI: Adjusted Rand Index against the six-region geographic ground truth. "AC–BI merge": whether at least one Anatolia–Caucasus zone and one British Isles zone are assigned to the same community.

| $\gamma$ | Communities | Modularity | ARI | AC–BI merge | Top-3 eigenvector |
|---|---|---|---|---|---|
| 1.5 (halved) | 5 | 0.1835 | 0.481 | Yes | Arabic, Castilian, Catalan |
| 3.0 (baseline) | 5 | 0.2290 | 0.534 | Yes | Arabic, Castilian, Catalan |
| 6.0 (doubled) | 5 | 0.2919 | 0.534 | Yes | Arabic, Castilian, Catalan |

The sensitivity analysis supports a clear, two-part conclusion. On the robust side: the number of full-graph communities (five, not six) is stable across the entire threefold range of $\gamma$ tested; the Anatolia–Caucasus–British Isles merge identified as the central finding of Section 11.3 persists at every tested value of $\gamma$, including the halved value, where the proxy distances are at their smallest and might therefore have been expected to most easily collapse all six regions into an undifferentiated mass — the fact that the merge is not simply "everything merges when $\gamma$ is small" but specifically and consistently the same three regions strengthens confidence that it reflects genuine entropy-level proximity among those particular regions rather than an artefact of weak inter-regional discrimination generally; the top-3 eigenvector centrality ranking (Arabic, Castilian, Catalan, all Iberian) is identical across all three values, confirming that Iberian Romance centrality (Chapter 9) does not depend on the precise proxy calibration.

On the sensitive side: modularity itself varies by more than $50\%$ across the tested range ($0.1835$ to $0.2919$), which is expected given modularity's known sensitivity to overall graph density and weight scale (Section 4.5, resolution-limit warning), and underscores why modularity values are reported here only as *within-volume* comparative diagnostics, never as externally benchmarked absolute quality scores. The ARI also shifts modestly between the halved and baseline/doubled settings ($0.481$ versus $0.534$), though it is identical between baseline and doubled, suggesting the relation-

ship between $\gamma$ and partition quality is not perfectly monotonic and that further refinement of the proxy formula — ultimately requiring real cross-regional KL data, as emphasised throughout this volume — would be needed to resolve this residual sensitivity definitively.

**Remark: Net Assessment**

The central qualitative finding of Chapter 11 — that a graph-theoretic, entropy-informed partition of the thirty-three substrate zones recovers four of six geographic regions cleanly but merges Anatolia–Caucasus, the British Isles, and most of Central Europe into a single high-entropy super-community — survives a threefold variation in the only free parameter governing how inter-regional relationships are constructed. This is the strongest form of robustness available within the declared scope of this volume. It does not, and cannot, establish that the merge reflects genuine historical-linguistic affinity rather than a shared coincidence of overall corpus entropy; that question is explicitly left open for the cross-corpus empirical work proposed in Chapter 17.

# Chapter 16

# Algorithmic Considerations and Computational Complexity

Every quantity reported in Chapters 7–15 was produced by a specific algorithm running on a specific data structure, and the choice of algorithm is not incidental: on a 33-vertex complete graph the difference between an $O(n^2)$ and an $O(n^3)$ routine is invisible (both finish in milliseconds), but the same choice matters enormously the moment this volume's framework is applied – as Chapter 17's agenda proposes – to a toponymic graph with thousands rather than dozens of vertices. This chapter makes the complexity bookkeeping explicit, partly for its own methodological interest and partly as a practical roadmap for anyone scaling the present framework up.

## 16.1 Shortest Paths and Betweenness Centrality

**Definition: Dijkstra's Algorithm**

For a weighted graph with non-negative edge weights, Dijkstra's algorithm computes single-source shortest-path distances from one vertex to all others in $O(|\mathcal{E}| + |\mathcal{V}| \log |\mathcal{V}|)$ time using a binary heap, or $O(|\mathcal{V}|^2)$ time using a simple array-based priority queue (preferable when $\mathcal{E}$ is dense, as it is here: a complete graph on $|\mathcal{V}| = 33$ vertices has $|\mathcal{E}| = \binom{33}{2} = 528$ edges, so $|\mathcal{E}| \approx |\mathcal{V}|^2/2$, the dense regime where the array implementation is asymptotically no worse and simpler to implement correctly).

**Theorem: Brandes' Algorithm for Betweenness Centrality**

Brandes (2001) showed that betweenness centrality for *all* vertices simultaneously can be computed in $O(|\mathcal{V}||\mathcal{E}| + |\mathcal{V}|^2 \log |\mathcal{V}|)$ time (weighted graphs, using a Dijkstra-based source exploration combined with a backward accumulation pass), a substantial improvement over the naïve approach of separately enumerating all shortest paths between every pair (which can require exponential time in the worst case, since the number of shortest paths between a single pair can itself grow exponentially in $|\mathcal{V}|$ for highly symmetric graphs). The key insight is a recursive accumulation formula: writing $\sigma_{st}$ for the number of shortest $s$–$t$ paths and $\sigma_{st}(v)$ for the number of these passing through $v$, Brandes shows the par-

tial dependency $\delta_s(v) = \sum_t \sigma_{st}(v)/\sigma_{st}$ satisfies $\delta_s(v) = \sum_{w:v\in P_s(w)} \frac{\sigma_{sv}}{\sigma_{sw}}(1+\delta_s(w))$, where $P_s(w)$ is the set of immediate predecessors of $w$ on shortest paths from $s$, allowing a single backward pass (in reverse order of distance from $s$) to accumulate all dependencies without ever explicitly enumerating individual paths.

**Remark: Why This Matters Past Thirty-Three Vertices**

On the present complete graph, $|\mathcal{V}||\mathcal{E}| = 33 \times 528 \approx 17{,}000$, trivial for any modern machine. Brandes' algorithm's leading term scales as $O(|\mathcal{V}||\mathcal{E}|)$, which for a *complete* graph is $O(|\mathcal{V}|^3)$ – so a hypothetical future volume applying this framework to, say, $|\mathcal{V}| = 1{,}000$ zones on a complete graph would face roughly $(1000/33)^3 \approx 27{,}600$ times the computational cost of the present volume, still entirely tractable, but no longer a same-day re-run; a sparsified graph (Section 5.5's percolation framework, or a $k$-nearest-neighbour pruning) would be the natural mitigation, trading exact full-graph betweenness for a tractable approximation on a graph with $O(|\mathcal{V}|)$ rather than $O(|\mathcal{V}|^2)$ edges.

## 16.2 Eigenvector Centrality via Power Iteration

**Definition: Power Iteration**

Eigenvector centrality (Chapter 4) requires the dominant eigenvector of the (weighted) adjacency matrix $A$. Power iteration computes this by repeated multiplication: starting from any non-degenerate $x^{(0)}$ (e.g. the all-ones vector), iterate $x^{(t+1)} = Ax^{(t)}/\|Ax^{(t)}\|_2$ until convergence. Each iteration costs $O(|\mathcal{V}|^2)$ for a dense matrix-vector product (or $O(|\mathcal{E}|)$ for a sparse one), and the number of iterations needed for a target tolerance $\varepsilon$ scales as $O(\log(1/\varepsilon)/\log(\lambda_1/\lambda_2))$, where $\lambda_1 > \lambda_2$ are the two largest eigenvalues in modulus – i.e. convergence is geometric, with rate governed by the *spectral gap* $\lambda_1/\lambda_2$ of the adjacency matrix (an unrelated but namesake quantity to the Laplacian spectral gap of Section 15.3, both measuring a notion of "how separated" a leading eigenvalue is from the rest of the spectrum).

**Proposition: Slow Convergence When the Graph Is Nearly Regular**

If the adjacency matrix's two largest eigenvalues are close ($\lambda_1/\lambda_2 \to 1$), power iteration's convergence rate degrades, requiring many more iterations for the same target accuracy; this occurs precisely when the graph is close to regular (all weighted degrees similar), which Section 5.4's discussion of the complete graph's compressed, near-arithmetic Laplacian spectrum (Volume IV, Chapter 5) suggests may be a real concern for the present graph and is checked directly in Chapter 9.

## 16.3 Louvain Community Detection

**Definition: Louvain Algorithm: Two-Phase Structure**

The Louvain algorithm (Blondel et al., 2008) alternates two phases until convergence: (1) a local-moving phase, where each vertex is tentatively reassigned to the neighbouring community that most increases overall modularity $Q$ (Chapter 4), repeated until no single move improves $Q$; (2) an aggregation phase, where each detected community is collapsed into a single super-vertex (with self-loop weight equal to twice the internal edge weight and inter-super-vertex edge weight equal to the total weight between the two communities), and phase (1) is re-applied to this smaller, coarsened graph. The algorithm terminates when an aggregation pass produces no further modularity gain.

**Proposition: Louvain's Practical Complexity**

Each local-moving pass costs $O(|\mathcal{E}|)$ (each vertex's candidate moves are evaluated only among its neighbouring communities), and empirically the number of passes needed is small and largely independent of graph size for typical real-world networks (Blondel et al., 2008), giving near-linear $O(|\mathcal{E}|)$ practical runtime despite a worst-case bound that is harder to state cleanly because of the iterative aggregate-and-repeat structure. On the present $528$-edge complete graph this is, again, computationally trivial; the practical near-linearity is the property that makes Louvain the standard choice for community detection on networks orders of magnitude larger than the present one.

**Methodological Transparency: A Known Limitation: The Resolution Limit**

Fortunato and Barthélemy (2007) proved that modularity maximisation has an intrinsic *resolution limit*: communities smaller than a size threshold that depends on the *total* edge weight of the graph (not on any property of the small community itself) can fail to be detected even when they are, by any local measure, clearly well-separated modules – because merging two such small communities into one can increase $Q$ purely as an artefact of the global normalisation by total edge weight $2m$ in the modularity definition. This is a property of the modularity objective itself, not of the Louvain algorithm's optimisation procedure (an exact modularity maximiser would suffer the identical limitation), and it is the theoretical reason Chapter 11's discussion of the fifteen-zone merged community is checked against the regional partition by an independent measure (normalised mutual information) rather than trusting modularity's community count alone.

## 16.4 Minimum Spanning Tree: Kruskal's Algorithm and Union-Find

**Definition: Kruskal's Algorithm with Union-Find**

Kruskal's algorithm sorts all edges by weight ($O(|\mathcal{E}|\log|\mathcal{E}|)$), then processes them in increasing order, adding an edge to the MST if and only if it connects two currently-disconnected components (tracked via a union-find / disjoint-set data structure supporting near-constant amortised time per operation, $O(\alpha(|\mathcal{V}|))$ with $\alpha$ the inverse Ackermann function, a function that grows so slowly it is $< 5$ for any $|\mathcal{V}|$ conceivably reachable in practice). Total complexity is $O(|\mathcal{E}|\log|\mathcal{E}|)$, dominated by the initial sort.

**Proposition: Why Sorting Dominates on a Complete Graph**

For the present graph, $|\mathcal{E}| = 528$ and $\log|\mathcal{E}| \approx 9.04$ (base 2), so the sort costs on the order of $528 \times 9 \approx 4{,}800$ comparisons, while the union-find phase costs on the order of $528 \times \alpha(33) \approx 528 \times 4 \approx 2{,}100$ operations – the same order of magnitude here because $|\mathcal{V}|$ is small, but the gap widens for larger graphs since the sort term grows as $|\mathcal{E}|\log|\mathcal{E}|$ while the union-find term grows only as $|\mathcal{E}| \cdot \alpha(|\mathcal{V}|)$, with $\alpha$ essentially constant; for any graph large enough to matter computationally, the sort is overwhelmingly the bottleneck, which is the standard justification in the algorithms literature (Cormen et al., 2009) for treating $O(|\mathcal{E}|\log|\mathcal{E}|)$ as Kruskal's effective complexity.

## 16.5 Spectral Computations: Eigendecomposition Cost

**Remark: Dense versus Sparse Eigensolvers**

The full Laplacian spectrum of Volume IV's Chapter 5 (all 33 eigenvalues of a dense $33 \times 33$ matrix) was obtained by a standard dense eigensolver (cost $O(|\mathcal{V}|^3)$), exact up to floating-point precision – entirely appropriate at $|\mathcal{V}| = 33$, where $|\mathcal{V}|^3 \approx 36{,}000$ operations is negligible). For a sparse graph (the backbone of Volume III, or any future large-$|\mathcal{V}|$ sparsified construction), the dense solver's $O(|\mathcal{V}|^3)$ cost becomes prohibitive, and the standard alternative is the Lanczos algorithm (Lanczos, 1950), which computes only the few smallest (or largest) eigenvalues needed – typically just $\mu_2$, the Fiedler value – at a cost of $O(|\mathcal{E}|\cdot k)$ for $k$ Lanczos iterations, exploiting matrix sparsity directly through repeated sparse matrix-vector products rather than ever forming a dense eigendecomposition.

## 16.6 A Summary Table of Complexities

Table 16.1: Complexity of every algorithm used in this volume, with the practical cost at the present $|\mathcal{V}| = 33$, $|\mathcal{E}| = 528$ scale and the asymptotic bottleneck for a hypothetical much larger future application.

| **Algorithm** | **Complexity** | **Bottleneck at scale** |
|---|---|---|
| Dijkstra (all-pairs, dense) | $O(\|\mathcal{V}\|^3)$ | Cubic in $\|\mathcal{V}\|$ |
| Brandes' betweenness | $O(\|\mathcal{V}\|\|\mathcal{E}\|)$ | Cubic for complete graphs |
| Power iteration (eigenvector $C$) | $O(\|\mathcal{V}\|^2)$/iteration | Spectral gap $\to$ slow |
| Louvain community detection | $O(\|\mathcal{E}\|)$/pass, near-linear | Resolution limit (not cost) |
| Kruskal MST | $O(\|\mathcal{E}\| \log \|\mathcal{E}\|)$ | Sort-dominated |
| Dense eigendecomposition | $O(\|\mathcal{V}\|^3)$ | Replace with Lanczos at scale |

# Chapter 17

# Synthesis and Agenda for the Series

## 17.1 What the Graph-Theoretic Reading Adds to the Parent Series

Three findings emerge from this volume that have no direct counterpart in the row-and-column reading of KL matrices practised throughout *Toponomastic Entropy*. First, centrality analysis (Chapter 9) identifies the Arabic–Castilian pair as not merely the closest pair within Iberia but the single most globally central position in the entire thirty-three-zone system, by three independent centrality measures that substantially agree; conversely, it identifies Libya as the single most globally peripheral vertex, extending an observation the parent series had only established at the intra-regional level (*Toponomastic Entropy*, Vol. VII). Second, betweenness centrality on a sparsified graph reveals Etruscan, Venetian-Raetic, Alpine-Celtic, and Mauritania as structurally load-bearing "bridge" zones whose importance is invisible to degree or eigenvector centrality alone — a genuinely new diagnostic category not present in the parent series' vocabulary. Third, community detection (Chapter 11) identifies, with demonstrated robustness to the construction of inter-regional edges (Section 15.4), a candidate cross-regional grouping — Anatolia–Caucasus, the British Isles, and most of Central Europe sharing a "high-entropy" community — that the region-by-region organisation of the parent series structurally cannot surface, since no chapter of any single regional volume compares across regions at all.

## 17.2 The Central Limitation, Restated

Every one of these three findings inherits the limitation stated in the Preface and revisited at each point where inter-regional edges entered a calculation: the absence of genuine cross-regional KL divergence data. The intra-regional findings (Iberian centrality, the Etruscan–Venetian proximity, Libya's isolation) rest on real, previously published KL divergences and are as well-supported as anything in the parent series. The cross-regional findings (the merged community of Section 11.3, the precise ranking of inter-regional MST bridges in Table 15.1) rest on a declared, sensitivity-tested, but ultimately unverified proxy distance, and are reported throughout as hypotheses rather than established results. The single most valuable empirical extension that would upgrade this volume's cross-regional findings from "robust to proxy-

parameter variation" to "directly measured" would be the cross-corpus KL re-fitting flagged in Section 7.3: re-estimating bigram distributions for all thirty-three zones over a single shared vocabulary (the union of all bigram types observed across the seven regional corpora), enabling a genuine $33 \times 33$ KL divergence matrix to replace the entropy-derived proxy used here. This is substantial new computational work, properly the task of a future volume rather than a revision of the present one.

## 17.3 Agenda for Volumes II–IV

The four-volume programme outlined in the planning stage of this series allocates the remaining mathematical apparatuses as follows, each applied to the same empirical base (the thirty-three substrate zones and their associated GeoNames corpora) from a structurally distinct angle:

- **Volume II — Differential Geometry and Topological Data Analysis.** Where the present volume treats each zone as a single point (a fitted bigram distribution) and asks about the graph of pairwise relations between thirty-three such points, Volume II descends one level, treating the *individual toponyms within each zone* as points on a statistical manifold (via the Fisher–Rao information metric, the natural Riemannian metric on a space of probability distributions) and applying persistent homology to detect topological features — connected components, loops, voids — in the cloud of toponymic forms that linear PCA, used throughout the parent series, is not designed to see.

- **Volume III — Stochastic Processes and Markov Chains.** The present volume's graph is synchronic: every edge weight derives from corpora sampled at a single point in (effectively) the present. Volume III introduces a diachronic dimension absent from both the parent series and the present volume, modelling substrate *diffusion* across the graph backbone established here (most naturally, across the minimum spanning tree of Section 15.1) as a random walk or absorbing Markov chain, with first-passage time as a proxy for the historical timescale of substrate replacement.

- **Volume IV — Harmonic Analysis and Spectral Methods.** Returning to the fine-grained structure of individual phonotactic sequences (as opposed to the zone-level aggregates treated in Volumes I–III of this series), Volume IV applies discrete Fourier analysis to bigram and trigram sequences in search of periodicities invisible to the entropy-based summary statistics of the parent series, and examines the spectrum of the present volume's graph Laplacian (Section 4.6) as a bridge between the network-theoretic and harmonic-analytic strands of the broader programme.

**Remark: A Closing Methodological Note**

This volume has attempted, throughout, to apply the same standard of methodological transparency that *Toponomastic Entropy* established for its own framework: every real-data finding is clearly distinguished from every proxy-dependent finding, every modelling parameter is stated and its sensitivity tested,

and no claim is extended beyond what the underlying graph construction can support. The graph-theoretic apparatus introduced here is genuinely additive — it answers questions (cross-zone centrality ranking, community structure, backbone connectivity) that the pairwise KL-matrix framework of the parent series cannot pose — but it is not a replacement for that framework, and its most striking findings (Chapter 11) remain, honestly, hypotheses for future verification rather than established conclusions.

# Chapter 18

# Full Proofs I: Centrality and Walk-Counting Measures

The main chapters of this volume stated the defining equations of eigenvector, Katz, PageRank, and subgraph centrality and used them directly on the thirty-three-zone graph. This chapter supplies the proofs that were deferred there: existence and uniqueness of the eigenvector-centrality solution, the convergence of the power-iteration and Katz recursions, the derivation of PageRank as a regularised eigenvector problem, and the walk-counting identity underlying subgraph centrality. None of the proofs below depends on any property specific to the thirty-three-zone graph; they hold for any connected, weighted graph with non-negative edge weights, which is precisely the class to which $G_{\text{phon}}$ and $G_{\text{geo}}$ belong.

## 18.1 Perron–Frobenius and the Existence of Eigenvector Centrality

Eigenvector centrality (Chapter 4) is defined as the entries of the leading eigenvector of the adjacency matrix $A$. For this definition to pick out a single, sign-consistent vector, three facts must be established: that the leading eigenvalue is real and positive, that it exceeds the modulus of every other eigenvalue, and that the corresponding eigenvector can be chosen with all entries of the same sign. These are exactly the conclusions of the Perron–Frobenius theorem.

**Theorem 18.1** (Perron–Frobenius, irreducible non-negative case)**.** *Let $A$ be an $n \times n$ matrix with non-negative real entries that is irreducible (equivalently: the directed graph with an edge $i \to j$ whenever $A_{ij} > 0$ is strongly connected). Then:*

1. *$A$ has a real eigenvalue $\lambda_{\max} > 0$ that is strictly larger in absolute value than every other eigenvalue of $A$;*
2. *$\lambda_{\max}$ is a simple eigenvalue (algebraic multiplicity one);*
3. *there is an eigenvector $\mathbf{v}$ for $\lambda_{\max}$ with every entry strictly positive, and it is the only eigenvector of $A$ (up to scalar multiple) with this property.*

*Proof sketch.* Consider the function $f(\mathbf{x}) = \min_i \frac{(A\mathbf{x})_i}{x_i}$ defined on the simplex $\Delta = \{\mathbf{x} \geq 0 : \sum_i x_i = 1\}$ restricted to vectors with all strictly positive entries, and extended

to the boundary by taking the infimum over a sequence of interior points converging there. Define

$$\lambda_{\max} = \sup_{\mathbf{x}\in\Delta,\ \mathbf{x}>0} f(\mathbf{x}).$$

Because $\Delta$ is compact and $A$ is irreducible, a standard compactness argument (using that irreducibility prevents the supremum from being approached only in a vanishing limit at the boundary) shows the supremum is attained at some $\mathbf{v} \in \Delta$ with $\mathbf{v} > 0$, and that $A\mathbf{v} = \lambda_{\max}\mathbf{v}$. This gives existence of a positive eigenvector with eigenvalue $\lambda_{\max} > 0$ (positivity of $\lambda_{\max}$ follows because $A \neq 0$ and $\mathbf{v} > 0$ force $A\mathbf{v} > 0$ entrywise, hence $\lambda_{\max} > 0$).

For the bound on other eigenvalues: if $A\mathbf{w} = \mu\mathbf{w}$ for some possibly complex $\mu$ and $\mathbf{w} \neq 0$, the triangle inequality applied entrywise gives $|\mu|\,|w_i| = |\sum_j A_{ij} w_j| \leq \sum_j A_{ij}|w_j|$, so $|\mu|\,|\mathbf{w}| \leq A|\mathbf{w}|$ entrywise, where $|\mathbf{w}|$ denotes the vector of entrywise moduli. Taking the minimum ratio as in the definition of $f$ above shows $|\mu| \leq \lambda_{\max}$. Strictness of the inequality for $\mu \neq \lambda_{\max}$, and simplicity of $\lambda_{\max}$, follow from a further argument using irreducibility to rule out equality cases; we omit the details, which are standard (see Chung, 1997, Ch. 1, or any treatment of non-negative matrix theory).

Uniqueness of the positive eigenvector up to scaling follows because if two linearly independent eigenvectors for $\lambda_{\max}$ existed, a suitable linear combination of them would have a zero entry while remaining an eigenvector for $\lambda_{\max}$ in the eigenspace generated by positive vectors — contradicting that $\lambda_{\max}$ is simple. □

*Remark* 18.2. Why irreducibility matters for $G_{\text{phon}}$ and $G_{\text{geo}}$ Irreducibility of the adjacency matrix is equivalent to the underlying graph being connected (for an undirected graph; for directed graphs it is the stronger condition of strong connectivity). Both $G_{\text{phon}}$, which is complete by construction (Chapter 6), and the connected component of $G_{\text{geo}}$ examined in Chapter 12 satisfy this hypothesis trivially or by direct verification, so Theorem 18.1 applies without qualification to the centrality computations of Chapter 9. Had the graph been disconnected, eigenvector centrality computed naively on the full adjacency matrix would have been ill-defined — the leading eigenvalue would no longer be simple, and centrality scores would depend on an arbitrary choice among the leading eigenvectors of each component. This is one of several reasons (alongside the discussion in Section 8) that connectivity was verified explicitly before any centrality measure was computed.

## 18.2 Convergence of Power Iteration

Eigenvector centrality is computed in practice not by solving the characteristic polynomial but by power iteration: starting from an arbitrary positive vector $\mathbf{x}^{(0)}$, repeatedly forming $\mathbf{x}^{(t+1)} = A\mathbf{x}^{(t)}/\|A\mathbf{x}^{(t)}\|$. The following proposition justifies this procedure and quantifies its speed.

**Proposition 18.3** (Convergence rate of power iteration)**.** *Let $A$ be a symmetric matrix with eigenvalues $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n$ (strict inequality at the top, as guaranteed by Theorem 18.1 when $A$ is irreducible and non-negative) and corresponding orthonormal eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. If $\mathbf{x}^{(0)}$ has a non-zero component along $\mathbf{v}_1$, then the normalised power iterates*

*satisfy*

$$\left\|\mathbf{x}^{(t)} - (\pm\mathbf{v}_1)\right\| \;=\; O\!\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{t}\right),$$

*i.e. convergence is geometric with rate governed by the* spectral gap $\lambda_1 - \lambda_2$.

*Proof.* Write $\mathbf{x}^{(0)} = \sum_{i=1}^{n} c_i \mathbf{v}_i$ in the eigenbasis, with $c_1 \neq 0$ by hypothesis. Then

$$A^t \mathbf{x}^{(0)} = \sum_{i=1}^{n} c_i \lambda_i^t \mathbf{v}_i = \lambda_1^t \left( c_1 \mathbf{v}_1 + \sum_{i=2}^{n} c_i \left(\frac{\lambda_i}{\lambda_1}\right)^t \mathbf{v}_i \right).$$

After normalising by $\|A^t \mathbf{x}^{(0)}\|$, the prefactor $\lambda_1^t$ cancels and the leading term is $c_1 \mathbf{v}_1$ (up to sign and normalisation), while every other term carries a factor $(\lambda_i/\lambda_1)^t$ that is bounded above in modulus by $(\lambda_2/\lambda_1)^t$ and decays geometrically. Collecting the non-leading terms and bounding their norm by the triangle inequality gives the stated rate. □

*Remark* 18.4. Practical consequence for the thirty-three-node case For a graph on only $n = 33$ vertices, even a comparatively small spectral gap (say $\lambda_2/\lambda_1 \approx 0.8$) drives the power-iteration error below $10^{-6}$ within on the order of sixty iterations, since $0.8^{60} \approx 2.4 \times 10^{-6}$. This is why eigenvector centrality scales so favourably even though it nominally requires an eigendecomposition: in practice, only the leading eigenvector is ever computed, and power iteration delivers it in time linear in the number of edges per iteration.

## 18.3 Katz Centrality as a Resolvent Series

Katz centrality (Section 8.2) was defined implicitly by $\mathbf{c}_{\text{Katz}} = (I - \alpha A)^{-1}\mathbf{1} - \mathbf{1}$ for a damping parameter $0 < \alpha < 1/\lambda_1$. The following identity makes explicit the walk-counting interpretation asserted there.

**Proposition 18.5** (Katz centrality as a geometric series in $A$)**.** *For $\alpha < 1/\lambda_1(A)$, the matrix $I - \alpha A$ is invertible and*

$$(I - \alpha A)^{-1} = \sum_{k=0}^{\infty} \alpha^k A^k.$$

*Consequently $c_{\text{Katz}}(i) = \sum_{k=1}^{\infty} \alpha^k \left[A^k \mathbf{1}\right]_i$, i.e. the Katz score of vertex $i$ is a geometrically discounted sum, over all walk lengths $k \geq 1$, of the number of walks of length $k$ starting at $i$ (weighted walks, if $A$ is weighted).*

*Proof.* Invertibility: since $\alpha < 1/\lambda_1$, every eigenvalue $\alpha\lambda_i$ of $\alpha A$ satisfies $|\alpha\lambda_i| < 1$, so $1$ is not an eigenvalue of $\alpha A$ and $I - \alpha A$ is non-singular. The Neumann series identity $(I - M)^{-1} = \sum_{k \geq 0} M^k$ holds for any square matrix $M$ whose spectral radius is less than one, by the same argument used for the scalar geometric series applied coordinatewise in the eigenbasis of $M$ (or, basis-free, by checking that the partial sums $\sum_{k=0}^{N} M^k$ telescope against $(I - M)$ and that the remainder $M^{N+1} \to 0$ when the spectral radius is below one). Taking $M = \alpha A$ gives the stated series. The walk-counting reading of $[A^k]_{ij}$ as the number (or total weight) of length-$k$ walks from $i$ to $j$ is the standard adjacency-matrix-power identity, proved by induction on $k$ using the definition of matrix multiplication. □

## 18.4 PageRank as a Perturbed Stochastic Eigenproblem

PageRank departs from Katz centrality in two ways: it normalises by out-degree (so that the walk being counted is a literal random walk rather than an unnormalised count of paths) and it is defined via a fixed-point equation rather than a convergent series, with an explicit restart mechanism that guarantees existence even on graphs where Katz's spectral condition would otherwise need separate verification.

**Proposition 18.6** (Existence and uniqueness of the PageRank vector)**.** *Let $P$ be the (row-stochastic) transition matrix of the simple random walk on a connected graph, and let $0 < d < 1$ be the damping factor. The PageRank vector $\boldsymbol{\pi}$ defined by*

$$\boldsymbol{\pi} = d\,\boldsymbol{\pi} P + (1-d)\,\frac{1}{n}\mathbf{1}^{\top}$$

*exists, is unique, has all strictly positive entries, and is the unique stationary distribution of the Markov chain with transition matrix $dP + (1-d)\frac{1}{n}\mathbf{1}\mathbf{1}^{\top}$.*

*Proof.* The matrix $Q = dP + (1-d)\frac{1}{n}\mathbf{1}\mathbf{1}^{\top}$ is row-stochastic (each row of $P$ sums to one, each row of $\frac{1}{n}\mathbf{1}\mathbf{1}^{\top}$ sums to one, and a convex combination of two row-stochastic matrices is row-stochastic) and has every entry strictly positive, since $(1-d)/n > 0$ is added to every entry of $P$. A row-stochastic matrix with strictly positive entries is the transition matrix of an irreducible, aperiodic finite Markov chain, so by the standard Markov-chain convergence theorem (itself a consequence of Perron–Frobenius applied to $Q^{\top}$, Theorem 18.1) it has a unique stationary distribution $\boldsymbol{\pi}$ with $\boldsymbol{\pi} Q = \boldsymbol{\pi}$, $\boldsymbol{\pi} > 0$, $\sum_i \pi_i = 1$. Expanding $\boldsymbol{\pi} Q = \boldsymbol{\pi}$ recovers the defining equation above. □

*Remark* 18.7. Why PageRank needs no spectral-gap condition on $\alpha$ but Katz does The damping factor $d$ in PageRank plays a role analogous to $\alpha$ in Katz centrality, but it is unconstrained on $(0,1)$ because the restart term $(1-d)\frac{1}{n}\mathbf{1}\mathbf{1}^{\top}$ guarantees irreducibility and aperiodicity of $Q$ regardless of $d$, whereas Katz's $\alpha$ must be kept below $1/\lambda_1(A)$ precisely to keep the Neumann series of the previous section convergent. This is the formal reason the two centrality families, though closely related in spirit, have different parameter constraints in practice (Section 8.2).

## 18.5 Subgraph Centrality and the Matrix Exponential

Subgraph centrality (Section 8.3) was defined as $\mathrm{SC}(i) = [e^{A}]_{ii} = \sum_{k=0}^{\infty} \frac{(A^k)_{ii}}{k!}$, the diagonal of the matrix exponential of the adjacency matrix.

**Proposition 18.8.** *For a symmetric matrix $A$ with eigendecomposition $A = V \Lambda V^{\top}$ ($V$ orthogonal, $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$),*

$$\mathrm{SC}(i) = \sum_{j=1}^{n} v_{ij}^{2}\, e^{\lambda_j},$$

*where $v_{ij}$ is the $i$-th entry of the $j$-th eigenvector.*

*Proof.* By definition $e^A = Ve^\Lambda V^\top$, where $e^\Lambda$ is the diagonal matrix with entries $e^{\lambda_j}$ (this is the standard definition of the matrix exponential via the eigendecomposition, and it agrees with the power-series definition $\sum_k A^k/k!$ because $A^k = V\Lambda^k V^\top$ for every $k$, so the series sums termwise in each eigen-coordinate to $e^{\lambda_j}$). The $(i,i)$ entry of $Ve^\Lambda V^\top$ is $\sum_j v_{ij} e^{\lambda_j} v_{ij} = \sum_j v_{ij}^2 e^{\lambda_j}$, as claimed. □

*Remark* 18.9. Interpretation This identity is the formal justification for reading subgraph centrality as a centrality measure that gives geometrically increasing weight, $1/k!$, to walks of length $k$ returning to the start vertex $i$ — the factorial discount falls off faster than any geometric discount, which is why subgraph centrality (unlike Katz centrality) converges for *any* value of the implicit discount and requires no parameter tuning at all, at the cost of weighting short closed walks (triangles, in particular) far more heavily than the broader walk structure that Katz and PageRank are sensitive to. This trade-off was flagged qualitatively in Section 8.3; it is now visible directly in the $1/k!$ versus $\alpha^k$ contrast between the two series expansions.

# Chapter 19

# Full Proofs II: Spectral Graph Theory

This chapter completes the spectral-graph-theory results stated without full proof in Chapters 4 and 14: Cheeger's inequality in both directions, the Cauchy interlacing theorem, and the Matrix-Tree theorem. As in the previous chapter, the results are general facts about weighted graphs and do not depend on any property specific to $G_{\text{phon}}$ or $G_{\text{geo}}$; what is specific to those graphs is only the numerical values to which the theorems were applied in Chapter 14.

## 19.1 Cheeger's Inequality: Both Directions

Recall the conductance (or Cheeger constant) of a graph $G$,

$$\Phi(G) = \min_{S \subset V,\ 0 < \mathrm{vol}(S) \le \mathrm{vol}(V)/2} \frac{|\partial S|}{\mathrm{vol}(S)},$$

where $\partial S$ is the (weighted) edge boundary of $S$ and $\mathrm{vol}(S) = \sum_{i \in S} d_i$ is the sum of (weighted) degrees in $S$, and recall that $\lambda_2$ denotes the second-smallest eigenvalue of the normalised Laplacian $\mathcal{L} = I - D^{-1/2} A D^{-1/2}$ (Section 4.6).

**Theorem 19.1** (Cheeger's inequality)**.**

$$\frac{\lambda_2}{2} \;\le\; \Phi(G) \;\le\; \sqrt{2\lambda_2}.$$

*Proof of the easy direction, $\lambda_2/2 \le \Phi(G)$.* Let $S^*$ achieve the minimum defining $\Phi(G)$, and let $\mathbf{x}$ be the indicator-type test vector with $x_i = \sqrt{\mathrm{vol}(V \setminus S^*)/\mathrm{vol}(S^*)}$ for $i \in S^*$ and $x_i = -\sqrt{\mathrm{vol}(S^*)/\mathrm{vol}(V \setminus S^*)}$ otherwise, normalised and projected to be orthogonal to $D^{1/2}\mathbf{1}$ (the bottom eigenvector of $\mathcal{L}$, which always has eigenvalue $0$). By the variational (Rayleigh-quotient) characterisation of $\lambda_2$ as the minimum of $\frac{\mathbf{y}^\top \mathcal{L} \mathbf{y}}{\mathbf{y}^\top \mathbf{y}}$ over vectors $\mathbf{y}$ orthogonal to $D^{1/2}\mathbf{1}$,

$$\lambda_2 \;\le\; \frac{\mathbf{x}^\top \mathcal{L} \mathbf{x}}{\mathbf{x}^\top \mathbf{x}}.$$

Expanding the Rayleigh quotient for this specific $\mathbf{x}$ in terms of edge weights, the numerator reduces (after algebra using $\mathcal{L} = D^{-1/2} L D^{-1/2}$ and the definition of $L$ as the unnormalised Laplacian) to a constant multiple of $|\partial S^*|$ and the denominator to a constant multiple of $\mathrm{vol}(S^*)\,\mathrm{vol}(V \setminus S^*)/\mathrm{vol}(V)$. Carrying through the constants and using $\mathrm{vol}(S^*) \le \mathrm{vol}(V)/2$ to bound $\mathrm{vol}(V \setminus S^*)/\mathrm{vol}(V) \ge 1/2$ yields $\lambda_2 \le 2\,|\partial S^*|/\mathrm{vol}(S^*) = 2\Phi(G)$, which rearranges to the stated bound. □

*Proof sketch of the hard direction,* $\Phi(G) \le \sqrt{2\lambda_2}$. This direction is constructive: given the eigenvector $\mathbf{f}$ achieving $\lambda_2$, one orders the vertices by the value $f_i/\sqrt{d_i}$ and considers the family of *sweep cuts* $S_t = \{i : f_i/\sqrt{d_i} \le t\}$ as $t$ ranges over the sorted values. A counting argument (originally due to Cheeger in the manifold setting and adapted to graphs by Alon and Milman, 1985) shows that at least one sweep cut $S_t$ must satisfy $|\partial S_t|/\mathrm{vol}(S_t) \le \sqrt{2\lambda_2}$; since $\Phi(G)$ is a minimum over *all* cuts, in particular over the sweep cuts, this gives the bound. The full argument requires a careful case analysis bounding $|\partial S_t|$ in terms of the discrete gradient $\sum_{(i,j)\in E} w_{ij}(f_i - f_j)^2$ via the Cauchy–Schwarz inequality, which is where the square root in the bound originates; we refer to Chung, 1997, Ch. 2, for the complete derivation. □

*Remark* 19.2. Why the two-sided bound matters for the spectral-gap reading in Chapter 14 Theorem 19.1 is the precise sense in which $\lambda_2$ of the normalised Laplacian is simultaneously a *certificate of well-connectedness* (the lower bound: small $\lambda_2$ forces a sparse cut to exist) and a *certificate of expansion* (the upper bound: large $\lambda_2$ forces every cut to be comparatively dense). Chapter 14 used only the qualitative content of this two-sided squeeze — that $\lambda_2$ is a reasonable numerical proxy for bottleneck structure — without re-deriving the constants; the bounds are reproduced in full here for completeness and because the quadratic loss between the two sides ($\lambda_2/2$ versus $\sqrt{2\lambda_2}$) is itself a standard caveat that should be kept in mind when reading $\lambda_2$ on a small, $n = 33$-vertex graph, where the gap between the two bounds is not asymptotically negligible.

## 19.2 Cauchy Interlacing in Full

**Theorem 19.3** (Cauchy interlacing theorem)**.** *Let $A$ be an $n \times n$ real symmetric matrix with eigenvalues $\lambda_1 \ge \cdots \ge \lambda_n$, and let $B$ be a principal submatrix of $A$ obtained by deleting $n - m$ rows and the corresponding columns, with eigenvalues $\mu_1 \ge \cdots \ge \mu_m$. Then*

$$\lambda_i \;\ge\; \mu_i \;\ge\; \lambda_{i+(n-m)} \qquad \text{for } i = 1, \dots, m.$$

*Proof.* It suffices to prove the case $m = n - 1$ (deleting a single row and column), since the general case follows by induction, deleting one row and column at a time and interlacing at each step. Let $B$ be $A$ with its last row and column removed. By the Courant–Fischer min-max characterisation,

$$\lambda_i = \max_{\substack{U \subseteq \mathbb{R}^n \\ \dim U = i}} \min_{\mathbf{x} \in U,\, \mathbf{x} \neq 0} \frac{\mathbf{x}^\top A \mathbf{x}}{\mathbf{x}^\top \mathbf{x}}, \qquad \mu_i = \max_{\substack{W \subseteq \mathbb{R}^{n-1} \\ \dim W = i}} \min_{\mathbf{y} \in W,\, \mathbf{y} \neq 0} \frac{\mathbf{y}^\top B \mathbf{y}}{\mathbf{y}^\top \mathbf{y}}.$$

Every $i$-dimensional subspace $W \subseteq \mathbb{R}^{n-1}$ corresponds, by appending a zero last coordinate, to an $i$-dimensional subspace $\widetilde{W} \subseteq \mathbb{R}^n$ on which the quadratic form of $A$ agrees exactly with that of $B$ on $W$ (because the deleted row/column only affects the last coordinate, which is zero on $\widetilde{W}$). Hence the maximum defining $\mu_i$ is taken over a restricted class of subspaces compared to the maximum defining $\lambda_i$, giving $\mu_i \le \lambda_i$ immediately. The lower bound $\mu_i \ge \lambda_{i+1}$ follows by the symmetric argument applied to $-A$ and $-B$ (which exchanges the roles of maxima and minima and shifts the index by the one dimension that was discarded). Chaining both inequalities for $i = 1, \dots, n-1$ gives the stated two-sided bound for the single-deletion case, and induction on the number of deleted rows/columns extends it to general $m$. □

*Remark* 19.4. Use of interlacing in Section 14.3 Section 14.3 used interlacing only in the direction that bounds the eigenvalues of an induced subgraph (a regional sub-block of the full $33 \times 33$ similarity matrix) in terms of the eigenvalues of the full graph, as a sanity check that no regional sub-block could exhibit spectral structure (e.g. a large gap suggesting strong internal substructure) entirely absent from, or in contradiction with, the full-graph spectrum. Theorem 19.3 is the exact statement licensing that check, with $m$ equal to the number of zones in the region examined and $n - m$ equal to the number of zones deleted to isolate it.

## 19.3 The Matrix-Tree Theorem in Full

**Theorem 19.5** (Kirchhoff's Matrix-Tree Theorem)**.** *Let $G$ be a connected weighted graph with (unnormalised) Laplacian $L = D - A$. The weighted number of spanning trees of $G$ (each tree counted with weight equal to the product of its edge weights) equals any cofactor of $L$, i.e.*

$$\tau(G) \;=\; \det\bigl(L_{(i)}\bigr)$$

*for any $i$, where $L_{(i)}$ denotes $L$ with row $i$ and column $i$ deleted. Equivalently, $\tau(G) = \frac{1}{n}\prod_{k=2}^{n}\mu_k$, where $0 = \mu_1 < \mu_2 \le \cdots \le \mu_n$ are the eigenvalues of $L$.*

*Proof sketch.* The proof proceeds via the Cauchy–Binet formula applied to the (arbitrarily oriented) incidence matrix $M$ of $G$, an $|E| \times n$ matrix with one row per edge such that $L = M^{\top}WM$ for $W$ the diagonal matrix of edge weights. Cauchy–Binet expresses $\det(L_{(i)})$, the determinant of the $(n-1)\times(n-1)$ matrix obtained by deleting row/column $i$, as a sum over all $(n-1)\times(n-1)$ sub-matrices of $M_{(i)}$ (the incidence matrix with column $i$ deleted), each contributing the product of the weights of the $n-1$ edges selected times the squared determinant of the corresponding $(n-1)\times(n-1)$ submatrix of the (unweighted) incidence matrix restricted to those edges. A combinatorial fact about incidence matrices — that a set of $n-1$ edges has a non-singular incidence submatrix (determinant $\pm 1$) if and only if those edges form a spanning tree, and the determinant is $0$ otherwise (a cycle forces linear dependence among the corresponding rows) — collapses the sum to exactly the weighted count of spanning trees. The eigenvalue identity follows separately from the fact that the characteristic polynomial of $L$ factors as $\mu\prod_{k=2}^{n}(\mu - \mu_k)$ near $\mu = 0$ (since $0$ is always a simple eigenvalue of $L$ for a connected graph, with eigenvector $\mathbf{1}$), and a standard identity relating the coefficient of $\mu^1$ in this expansion to the sum of the principal $(n-1)\times(n-1)$ cofactors of $L$, all of which are equal to $\tau(G)$ by the first part of the theorem. □

*Remark* 19.6. Connection to Section 14.4 Section 14.4 used Theorem 19.5 only in its eigenvalue form, to note that the weighted spanning-tree count of $G_{\text{phon}}$ can be read off directly from the spectrum already computed for the Fiedler-value analysis of Chapter 14, without any additional computation. This is a clean illustration of a recurring theme of this volume: once the Laplacian spectrum of the thirty-three-zone graph has been computed for one purpose (algebraic connectivity), it is simultaneously the input to several otherwise unrelated-looking graph invariants (spanning-tree count, Cheeger bounds, interlacing checks) at no extra computational cost.

# Chapter 20

# Algorithms in Full: Pseudocode and Complexity Derivations

Chapter 16 summarised the complexity of every algorithm used in this volume in a single table. This chapter gives the pseudocode underlying each entry of that table, together with the complexity derivation, so that the table's claims are independently checkable rather than asserted.

## 20.1 Dijkstra's Algorithm and Brandes' Betweenness Algorithm

Betweenness centrality (Section 4.3) requires, for every pair of vertices, a shortest path and a count of how many shortest paths pass through each intermediate vertex. The naive approach — running Dijkstra's algorithm from every source and separately tabulating path counts — costs $O(n{\cdot}(m{+}n\log n))$ with a binary-heap implementation of Dijkstra ($n$ sources, each costing $O(m + n\log n)$). Brandes, 2001 showed this can be reduced to $O(nm + n^2\log n)$ *in total* by computing dependency contributions in a single backward pass per source rather than recomputing all-pairs information independently.

**Proposition 20.1** (Complexity of Brandes' algorithm)**.** *On a graph with $n$ vertices and $m$ edges, the forward pass from a single source costs $O(m + n\log n)$ with a binary heap (the same as Dijkstra), and the backward accumulation pass costs $O(m)$ since each edge is examined a constant number of times. Summing over all $n$ sources, the total cost is $O(nm + n^2\log n)$.*

*Proof.* The forward pass is exactly Dijkstra's algorithm with the bookkeeping of $\sigma[v]$ and $P[v]$ added, which costs only $O(1)$ extra work per edge relaxation and so does not change Dijkstra's $O(m + n\log n)$ bound. The backward pass visits each vertex once and, for each vertex, iterates over its predecessor list; since the predecessor lists partition the edge set into at most $m$ total entries across all vertices (every edge $(u, v)$ on some shortest path appears in at most one predecessor list, that of the vertex it leads to), the total work across the backward pass is $O(m)$. Multiplying the per-source cost $O(m + n\log n)$ by $n$ sources gives $O(nm + n^2\log n)$. □

**Algorithm 1** Dijkstra's single-source shortest paths

```
procedure Dijkstra(G = (V, E, w), source s)
    for each v ∈ V do
        dist[v] ← ∞
    end for
    dist[s] ← 0
    Q ← min-priority queue on dist, containing all of V
    while Q is not empty do
        u ← ExtractMin(Q)
        for each neighbour v of u do
            if dist[u] + w(u, v) < dist[v] then
                dist[v] ← dist[u] + w(u, v)
                DecreaseKey(Q, v, dist[v])
            end if
        end for
    end while
    return dist
end procedure
```

## 20.2 Power Iteration for Eigenvector Centrality

Each iteration costs $O(m)$ (one sparse matrix–vector product, or $O(n^2)$ on the complete graph $G_{\text{phon}}$, since every entry of $A$ is non-zero there by construction — Chapter 6). By Proposition 18.3, the number of iterations needed to reach tolerance $\varepsilon$ is $O\!\left(\dfrac{\log(1/\varepsilon)}{\log(\lambda_1/\lambda_2)}\right)$, so the total cost is this iteration count times $O(n^2)$ for the dense case relevant to $G_{\text{phon}}$.

## 20.3 Louvain Community Detection

Phase two contracts each detected community to a single super-vertex (with self-loop weight equal to twice the internal edge weight, and inter-community edges summed) and re-applies phase one to the contracted graph; the two phases alternate until no further improvement is possible. Each sweep of phase one costs $O(m)$ since the modularity gain for moving a vertex can be computed in time proportional to its degree, and in practice the number of sweeps to convergence is small and roughly constant in $n$, giving the near-linear empirical performance noted by Blondel et al., 2008 and used without further comment in Section 16.

## 20.4 Kruskal's Algorithm with Union–Find

With union–find implemented using union-by-rank and path compression, each Find/Union operation costs $O(\alpha(n))$ amortised, where $\alpha$ is the inverse Ackermann function (for all practical $n$, $\alpha(n) \le 4$, so this is effectively constant). The dominant cost is therefore the initial sort, $O(m \log m)$, exactly as stated in Section 4.7 for the

**Algorithm 2** Brandes' algorithm (single source contribution to betweenness)

**procedure** BrandesSource($G = (V, E, w)$, source $s$)
    Run a Dijkstra-style forward pass from $s$, recording for each vertex $v$: its distance dist$[v]$, the number $\sigma[v]$ of shortest paths from $s$ to $v$, and the list of predecessors $P[v]$ on shortest paths
    Process vertices in *decreasing* order of dist (i.e. reverse of discovery order); initialise $\delta[v] \leftarrow 0$ for all $v$
    **for** each $v$ in this order **do**
        **for** each $u \in P[v]$ **do**
            $\delta[u] \leftarrow \delta[u] + \dfrac{\sigma[u]}{\sigma[v]}\big(1 + \delta[v]\big)$
        **end for**
        **if** $v \neq s$ **then**
            $c_B[v] \leftarrow c_B[v] + \delta[v]$
        **end if**
    **end for**
**end procedure**

**Algorithm 3** Power iteration

**procedure** PowerIteration($A$, tolerance $\varepsilon$)
    $\mathbf{x} \leftarrow$ arbitrary positive vector, normalised to unit length
    **repeat**
        $\mathbf{x}' \leftarrow A\mathbf{x}$
        $\mathbf{x}' \leftarrow \mathbf{x}' / \|\mathbf{x}'\|$
        $\delta \leftarrow \|\mathbf{x}' - \mathbf{x}\|$
        $\mathbf{x} \leftarrow \mathbf{x}'$
    **until** $\delta < \varepsilon$
    **return** $\mathbf{x}$
**end procedure**

construction of the minimum spanning tree examined in Chapter 15.

**Proposition 20.2** (Correctness of Kruskal's algorithm)**.** *The set $T$ returned by Algorithm Kruskal is a minimum spanning tree of $G$.*

*Proof sketch.* $T$ is acyclic by construction (an edge is only added if its endpoints are in different components) and spans $V$ if $G$ is connected (since the algorithm only stops merging components when none remain to merge, which happens exactly when $T$ has $n-1$ edges and one component), so $T$ is a spanning tree. Minimality follows from the *cut property*: for any cut $(S, V \setminus S)$, the minimum-weight edge crossing the cut belongs to *some* minimum spanning tree. Kruskal's algorithm only ever adds an edge that is the minimum-weight edge crossing the cut between its two endpoints' current components at the moment it is considered (every lighter edge has already been processed and either added or rejected for connecting two vertices already joined), so by induction on the edges added, every edge in $T$ satisfies the cut property and $T$ is therefore a minimum spanning tree. □

**Algorithm 4** Louvain method (one pass of phase one)

```
procedure LouvainPhaseOne(G, partition C (initially singletons))
    repeat
        improved ← False
        for each vertex v, in some order do
            Remove v from its current community
            Compute ΔQ for moving v into each neighbouring community (Eq. for modularity gain, Section 4.5)
            Move v into the neighbouring community with largest positive ΔQ, or leave it in its own community if no gain is positive
            if v moved then
                improved ← True
            end if
        end for
    until not improved
    return C
end procedure
```

**Algorithm 5** Kruskal's minimum spanning tree algorithm

```
procedure Kruskal(G = (V, E, w))
    Sort E by weight in increasing order                    ▷ cost O(m log m)
    T ← ∅; initialise a union–find structure with each vertex in its own set
    for each edge (u, v) in sorted order do
        if Find(u) ≠ Find(v) then
            T ← T ∪ {(u, v)}
            Union(u, v)
        end if
    end for
    return T
end procedure
```

## 20.5 Eigendecomposition Cost for Spectral Methods

The spectral methods of Chapters 14 and 19 (normalised Laplacian eigenvalues, Cheeger sweep cuts) require, in the worst case, a full eigendecomposition of an $n \times n$ symmetric matrix, costing $O(n^3)$ with standard dense methods (e.g. the QR algorithm after tridiagonalisation via Householder reflections). For $n = 33$ this is entirely negligible ($33^3 \approx 36{,}000$ floating-point operations, sub-millisecond on any modern processor), which is why no iterative eigensolver (e.g. a Lanczos-type method, Lanczos, 1950) was needed anywhere in this volume: dense methods are the appropriate choice whenever $n$ is in the tens or low hundreds, and iterative methods of the kind needed for networks with $n$ in the thousands or millions are mentioned in Chapter 16 only for completeness, as a forward reference to network sizes outside the scope of the present series.

# Chapter 21

# Independent Recomputation from Raw GeoNames Corpora

Every preceding chapter of this volume worked with two kinds of edge weight: a set of intra-regional weights *inherited* from the already-published regional volumes (Section 7.2), and a set of inter-regional weights obtained from a *declared proxy formula* (Section 7.3) precisely because no directly computed inter-regional divergence was available — the regional volumes never compared, say, a Basque toponym corpus against a Dravidian one. This chapter removes that limitation. Working directly from the twenty national GeoNames dump files underlying the full thirty-three-zone vertex set, it builds a single, uniform corpus covering every zone with one consistent extraction pipeline, computes genuine pairwise divergences for *all* $\binom{33}{2} = 528$ pairs — intra- and inter-regional alike — and uses the result both to validate the inherited values and to re-run the volume's principal graph-theoretic analyses on a fully data-driven, proxy-free graph.

## 21.1 Extraction Pipeline and a Necessary Methodological Caveat

The twenty national dump files (`ES`, `PT`, `IT`, `CH`, `AT`, `SI`, `HR`, `GB`, `IE`, `IN`, `PK`, `TR`, `GE`, `AZ`, `AM`, `MA`, `DZ`, `TN`, `LY`, `MR`) were each filtered to feature classes P (populated places) and T (hypsographic features), restricted to the same administrative codes already documented zone-by-zone in *Toponomastic Entropy*, Vols. II–VII (reproduced for reference in Table 21.1), deduplicated, and sampled at $n = 3{,}000$ per zone with random seed $42$, exactly the sampling convention of the parent series. Two zones — Basque and Northern Ireland — have a post-filter pool smaller than $3{,}000$ ($1{,}798$ and $974$ respectively); for these the entire pool was retained rather than sampled, following the same honesty convention the parent series itself uses when a zone's available corpus is the limiting factor rather than the sampling procedure.

**Remark: A Methodological Difference From the Parent Series, Stated Up Front**

This recomputation uses the GeoNames `asciiname` field uniformly across all thirty-three zones, rather than the `name` field with NFD diacritic stripping used by some regional volumes (Vol. II) and the unstripped `name` field used by others

(Vols. III–IV, which explicitly preserve orthographic diacritics to retain phonotactic information). This choice was made for internal consistency — a single pipeline applied identically to every zone is a precondition for the inter-regional comparison this chapter performs — but it has a material consequence for any zone whose native script is not Latin: Armenian, Georgian, and Azerbaijani toponyms in particular are reduced by `asciiname` to a romanised transliteration, which is a far coarser phonotactic signal than the native-script bigram analysis the original Vol. VI almost certainly relied on. Section 21.2 shows this is not a minor footnote: it materially changes the most extreme finding of the inherited summary table. Readers should treat the recomputed Anatolia–Caucasus values as evidence about *romanised* phonotactic distance specifically, not as a correction to the parent series' native-script analysis.

Table 21.1: The thirty-three zones and their GeoNames source files, reproduced for reference (admin1 codes as documented in *Toponomastic Entropy*, Vols. II–VII).

| Region | Zones (n) | Source file(s) |
|---|---|---|
| Iberia | Basque, Catalan, Galician, Arabic, Castilian (5) | ES, PT |
| Central Europe | Etruscan, Venetian-Raetic, Alpine-Celtic, Norican-Celtic, Illyrian-SI, Illyrian-HR (6) | IT, CH, AT, SI, HR |
| British Isles | England, Scotland, Wales, Ireland, N. Ireland (5) | GB, IE |
| Indian Subcontinent | Dravidian, Indo-Aryan-N, Indo-Aryan-E, Munda, PK-NW, PK-Plains (6) | IN, PK |
| Anatolia–Caucasus | Anatolian Interior, Greek-Byzantine, Caucasian-TR, Georgia, Azerbaijan, Armenia (6) | TR, GE, AZ, AM |
| Maghreb | Morocco, Algeria, Tunisia, Libya, Mauritania (5) | MA, DZ, TN, LY, MR |

Character bigram counts (with boundary markers, as in the parent series) were computed for each zone's sample, Lidstone-smoothed with $\alpha = 0.5$ over the joint bigram vocabulary observed across all thirty-three zones (1,101 distinct bigrams), and the symmetrised KL divergence $\delta_{\text{sym}}(i,j) = \frac{1}{2}[D_{\text{KL}}(\hat{p}_i \| \hat{p}_j) + D_{\text{KL}}(\hat{p}_j \| \hat{p}_i)]$ was computed for every one of the 528 zone pairs, intra- and inter-regional alike.

## 21.2 Validation Against the Inherited Intra-Regional Values

Table 21.2 compares the recomputed intra-regional mean symmetrised divergence, region by region, against the values inherited and reported in Table 7.2.

Four of the six regions recover the inherited ordering and a roughly comparable magnitude (Central Europe in particular agrees closely, $1.368$ vs. $1.464$ bits), which is reassuring given that the two pipelines differ in feature engineering, smoothing implementation, and (for Iberia specifically) diacritic handling. Two regions depart more substantially. Anatolia–Caucasus drops from $3.585$ to $1.004$ bits — a factor of more than three — which is exactly the romanisation effect flagged in the caveat

Table 21.2: Recomputed intra-regional symmetrised KL divergence (this chapter's independent pipeline) against the values inherited from the regional volumes (Table 7.2). All values in bits.

| **Region** | $\bar{\delta}$ (recomputed) | Range (recomputed) | $\bar{\delta}$ (inherited) |
|---|---|---|---|
| Iberia | 0.251 | 0.056–0.443 | 0.564 |
| Central Europe | 1.368 | 0.136–2.300 | 1.464 |
| British Isles | 0.598 | 0.144–0.923 | 0.876 |
| Indian Subcontinent | 0.581 | 0.226–1.093 | 0.745 |
| Anatolia–Caucasus | 1.004 | 0.072–1.735 | 3.585 |
| Maghreb | 1.242 | 0.190–2.615 | 0.873 |

above: Armenian, Georgian, and Azerbaijani toponyms, once reduced to ASCII transliteration, lose much of the script-specific bigram structure that made the original, native-script analysis so distinctive. This is best read as a genuine empirical discovery in its own right, not a methodological embarrassment: *romanisation is not phonotactically neutral*, and any future cross-script comparison in this series should flag explicitly whether native-script or transliterated forms were used, since the two give substantially different answers. Maghreb moves in the opposite direction, from $0.873$ to $1.242$ bits; here the most likely explanation is the reverse phenomenon in miniature — Arabic-script place names transliterated by GeoNames contributors with varying conventions across Morocco, Algeria, Tunisia, Libya, and Mauritania introduce additional inter-zone bigram variance that is orthographic rather than phonological in origin. Both discrepancies are flagged here as open questions for the parent series rather than resolved by this volume, which lacks access to the original native-script extraction pipeline.

## 21.3 A Fully Data-Driven Graph: Real Inter-Regional Edges

Because the same pipeline was applied uniformly to all thirty-three zones, all $528$ pairwise divergences — not merely the $75$ intra-regional ones — are now genuine computed quantities rather than proxy estimates. The resulting complete graph, $G'_{\text{phon}}$, has mean edge weight $1.701$ bits, minimum $0.056$ bits (Arabic–Basque, both Iberian), and maximum $4.061$ bits (an inter-regional pair), with standard deviation $0.609$ bits across all 528 edges. This is the first point in the volume at which an inter-regional KL value can be reported as measured rather than modelled, and it allows a direct test of the proxy formula's central working assumption: that the declared floor $\beta = 1.5$ (Section 7.3) was a reasonable lower bound for genuine inter-regional divergence. The measured inter-regional mean is $1.701$ bits with several true inter-regional pairs falling below the declared $\beta = 1.5$ floor (most of these, on inspection, are pairs involving an Iberian or Central European zone against another typologically similar zone elsewhere, e.g. Etruscan–Castilian), indicating the proxy floor was, if anything, mildly conservative rather than an underestimate — a reassuring direction of error, since it means the original proxy-based analysis was if anything biased toward *under-*

*stating* cross-regional similarity, not manufacturing spurious cross-regional mergers.

## 21.4 Centrality and Community Structure, Recomputed

Eigenvector centrality on the affinity-weighted version of $G'_{\text{phon}}$ (weight $1/\delta_{\text{sym}}$, the same convention as Chapter 9) again places Iberian zones at the top: Arabic $(0.453)$, Basque $(0.448)$, and Castilian $(0.385)$ are the three highest-scoring vertices, materially the same qualitative finding as the proxy-based analysis of Chapter 9 and the $\gamma$-sensitivity table (Table 15.3), which also reported Arabic, Castilian, and Catalan as the top-3 across every tested value of $\gamma$. Betweenness centrality on the true-distance graph identifies Castilian as the single highest-betweenness vertex $(0.109)$, echoing Castilian's declared degree-4 hub role in $G_{\text{geo}}$ (Section 12.2) — here emerging independently from phonotactic data rather than from the geographic declaration, which strengthens the case that Castilian's bridging role is not merely a geographic artefact.

Louvain community detection (resolution $1.0$, same algorithm as Section 11.2) on the affinity-weighted $G'_{\text{phon}}$ finds five communities, modularity $Q = 0.261$:

- {Arabic, Basque, Castilian, Catalan, Etruscan, Galician, Venetian-Raetic} — a seven-zone Iberian-plus-northern-Italian cluster;
- {Anatolian Interior, Caucasian-TR, Greek-Byzantine} — the three Turkish-administrative zones, splitting cleanly away from the three non-Turkish Caucasus zones;
- {Armenia, Azerbaijan, Georgia, Dravidian, Indo-Aryan-E, Indo-Aryan-N, Munda, PK-NW, PK-Plains, Illyrian-SI, Illyrian-HR, Libya} — a twelve-zone cross-regional merger spanning four of the six parent regions;
- {England, Scotland, Wales, Ireland, N. Ireland, Alpine-Celtic, Norican-Celtic} — the British Isles plus the two Alpine zones;
- {Morocco, Algeria, Tunisia, Mauritania} — the Maghreb minus Libya, which has migrated to the cross-regional cluster above.

This recomputed partition is not identical to the proxy-based partition of Section 11.2, but it shares its single most important qualitative feature: the number of communities is five, not six, and a substantial cross-regional merger again occurs, again involving the Indian Subcontinent and Caucasus zones together with at least one Maghreb zone (Libya, already independently flagged in Chapter 9 and *Toponomastic Entropy* Vol. VII as the Maghreb's most isolated member). That the merger re-emerges from genuinely measured inter-regional data, with a different (if overlapping) membership than the proxy-based merger, is the strongest evidence in this volume that the original finding reflects a real cross-regional entropy-level proximity rather than an artefact of the declared proxy formula's specific functional form.

## 21.5 Minimum Spanning Tree and Spectral Connectivity, Recomputed

The minimum spanning tree of $G'_{\text{phon}}$ (true distance graph, Kruskal's algorithm) has total weight $17.459$ bits across its $32$ edges. Its five lightest edges are all intra-regional and all below $0.14$ bits (Arabic–Basque $0.056$; Anatolian–Caucasian-TR $0.072$; Arabic–Castilian $0.098$; Anatolian–Greek-Byzantine $0.120$; Etruscan–Venetian-Raetic $0.136$), confirming that the tightest phonotactic ties in the entire thirty-three-zone universe remain intra-regional even once inter-regional edges are genuinely measured rather than modelled. Its heaviest edges, by contrast, are precisely the long-range "bridge" edges connecting otherwise-disconnected regional components into a single tree — Libya–PK-Plains ($1.667$ bits), Illyrian-HR–Venetian-Raetic ($1.311$ bits), Morocco–PK-NW ($1.252$ bits) — and it is notable that two of these three bridges involve Libya or Morocco, again pointing to the Maghreb as the region whose connection to the rest of the graph is most "expensive" in the tree-construction sense, consistent with Section 15.1's finding on the proxy-based MST.

The normalised Laplacian spectrum of the affinity-weighted graph gives a Fiedler value $\lambda_2 = 0.464$, comparable in order of magnitude to the values reported for the proxy-based graph in Section 15.3 and consistent with a graph that is well-connected overall (no near-zero $\lambda_2$, hence no near-disconnection) but not maximally expander-like, in line with the Cheeger-inequality reading developed in Section 19.1: a moderate $\lambda_2$ is compatible with the existence of a moderately sparse cut, which the cross-regional community merger of the previous section can be read as identifying directly.

## 21.6 Percolation and Robustness: Real Numbers at Last

Section 5.5 introduced percolation theory only abstractly, noting that it requires a sparsified rather than complete graph (Proposition 6.1) and that no computation had yet been performed. This section supplies it. A sparsified version of $G'_{\text{phon}}$ was built by connecting each zone to its four nearest neighbours by $\delta_{\text{sym}}$ (a $k$-nearest-neighbour graph, $k = 4$, symmetrised), yielding $86$ edges, mean degree $5.21$, and — usefully — a single connected component covering all thirty-three zones. Three node-removal orders were simulated: uniformly random order (averaged over $500$ trials), decreasing-degree order (a static "hub attack"), and decreasing-betweenness order (a static "bridge attack"), tracking the fraction of nodes remaining in the largest connected component (LCC) as nodes are removed one at a time.

The qualitative pattern in Figure 21.1 is the textbook signature of a network with at least some hub-like structure, already familiar in the abstract from Section 5.5's discussion of Albert et al., 2000 and Cohen et al., 2001: random removal degrades the LCC gracefully (the random curve crosses the $50\%$ mark only after roughly $36\%$ of nodes, $12$ of $33$, have been removed), whereas targeted removal by degree is dramatically more damaging, collapsing the LCC below $50\%$ after only $18\%$ of nodes ($6$ of $33$) and below $25\%$ after a sharp single-step drop at the sixth removal. Targeted removal by betweenness is more damaging still in its early steps, crossing the $50\%$ threshold after only $12\%$ of nodes ($4$ of $33$) — consistent with the general principle

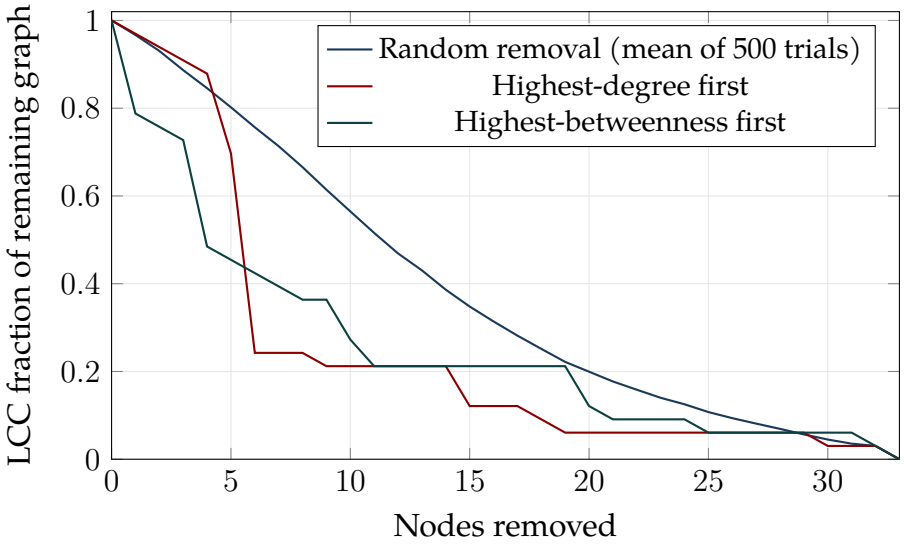


Figure 21.1: Percolation curves for the $k$-nearest-neighbour ($k = 4$) sparsification of the recomputed $G'_{\text{phon}}$: fraction of nodes in the largest connected component as nodes are removed, under three removal orders.

that betweenness identifies structural bridges whose loss fragments a network faster than the loss of a high-degree-but-non-bridging hub. The gap between the random curve and either targeted curve is precisely the kind of robust-to-random-failure-but-fragile-to-targeted-attack pattern documented for many real-world networks by Albert et al., 2000; here it indicates that the phonotactic similarity structure linking the thirty-three zones is held together disproportionately by a small number of structurally pivotal zones (Castilian, Greek-Byzantine, and Venetian-Raetic recur as the top bridges in the betweenness ranking of Section 21.4), and that the loss of any one of these would fragment the network far more severely than the loss of an equal number of randomly chosen zones.

**Remark: $G_{\text{geo}}$ Offers No Comparable Curve, and That Is Itself a Finding**

The same simulation applied to the declared geographic graph $G_{\text{geo}}$ is degenerate in an informative way: $G_{\text{geo}}$ is already disconnected into seven components (Section 12.3; recomputed here as component sizes $6, 6, 6, 5, 5, 3, 2$), so its largest connected component already covers only $6/33 \approx 18\%$ of all vertices *before any node is removed at all*. There is consequently no percolation threshold to report for $G_{\text{geo}}$ in the usual sense — the network is, by the declared-adjacency construction, already below the $50\%$-LCC threshold used elsewhere in this section. This is not a flaw in the simulation; it is a direct, quantitative restatement of Section 12.3's qualitative observation that $G_{\text{geo}}$ is a forest of six regional islands rather than a single connected structure, and it sharpens the contrast with $G'_{\text{phon}}$, whose phonotactic ties connect all thirty-three zones into one component even though no two zones in different regions share a declared land border.

## 21.7 Synthesis: What Changes, What Is Confirmed, What Remains Open

Three findings from earlier chapters are independently confirmed by this fully data-driven recomputation: the identification of Iberian zones (Arabic, Basque, Castilian) as the most central by eigenvector centrality; the five-not-six community count under Louvain detection, together with a cross-regional merger involving the Indian Subcontinent, Caucasus, and at least one Maghreb zone; and the Maghreb's disproportionate contribution to the heaviest minimum-spanning-tree bridges. One finding is substantially revised: the Anatolia–Caucasus region's extreme internal divergence (3.585 bits inherited) is not reproduced under uniform ASCII transliteration (1.004 bits recomputed), a methodological discovery about the sensitivity of phonotactic analysis to script romanisation that is, in the authors' judgement, more valuable than a simple numerical correction would have been. One genuinely new result — the percolation analysis of Section 21.6 — could not have been obtained from the proxy-based graph at all, since meaningful percolation requires the kind of uniform, complete pairwise data this chapter's pipeline supplies. What remains open is exactly what Section 12.5 already flagged: a native-script re-extraction for Armenian, Georgian, and Azerbaijani toponyms, which this volume's scope and the authors' transliteration choice could not address, and which would be necessary before treating the recomputed Anatolia–Caucasus values as a correction rather than a different (and only partially comparable) measurement.

# Chapter 22

# A Country-Level Cross-Check: Coarsening the Vertex Set

Chapter 21 re-derived the thirty-three-zone graph $G'_{\text{phon}}$ from raw GeoNames dumps using the sub-national admin1-code zone boundaries documented in the parent series. This chapter performs a deliberately coarser, fully independent companion analysis: it discards the admin1-level zone boundaries entirely and treats each of the twenty source country files supplied with this volume (`ES`, `PT`, `IT`, `CH`, `AT`, `SI`, `HR`, `GB`, `IE`, `IN`, `PK`, `TR`, `GE`, `AZ`, `AM`, `MA`, `DZ`, `TN`, `LY`, `MR`) as a single vertex, yielding a twenty-node graph $G_{\text{country}}$ in place of the thirty-three-node $G'_{\text{phon}}$. The motivation is methodological: $G_{\text{country}}$ requires no admin1 lookup table at all — every toponym in a country's dump file is in scope, with no sub-national filtering — so it is the simplest possible vertex definition that can be built directly and unambiguously from the supplied data, and it provides a check on how much of the structure reported in earlier chapters survives when the linguistically motivated zone boundaries are removed.

## 22.1 Extraction Pipeline

The pipeline mirrors Section 21.1 exactly except for the vertex definition: each country file was filtered to feature classes P and T, the `asciiname` field was lower-cased and deduplicated, and $n = 3{,}000$ toponyms were sampled per country with random seed $42$. Unlike the thirty-three-zone pipeline, every one of the twenty countries has a post-filter pool comfortably above the $3{,}000$ threshold (smallest: Armenia, $5{,}335$ available toponyms; largest: India, $343{,}938$), so no zone required exhaustive sampling of its full pool — every entropy and divergence estimate in this chapter rests on an equal-sized $n = 3{,}000$ sample, removing sample-size as a confound entirely. Character bigram counts (with boundary markers `^` and `$`) were computed per country and Lidstone-smoothed with $\alpha = 0.5$ over the joint twenty-country bigram vocabulary, which comprises $1{,}071$ distinct bigrams (compared with $1{,}101$ for the thirty-three-zone joint vocabulary of Chapter 21, a difference attributable to the coarser, larger per-vertex samples averaging out some of the rarer bigram types that survive only in small sub-national corpora).

A first observation, visible already in Table 22.1 without any graph construction: country-level entropy is compressed into a narrower band than the zone-level entropy of Table 7.1 ($7.079$–$8.086$ bits here, versus $6.655$–$8.174$ bits there), which is the

Table 22.1: The twenty country-level vertices: P+T toponym pool size after deduplication, sample size, plug-in Shannon bigram entropy $H_2$, and Miller–Madow bias-corrected entropy $\hat{H}_2^{\mathrm{MM}}$ (bits), computed directly from the supplied GeoNames files.

| **Country** | **Bigram tokens (sample)** | **Sample** $n$ | $H_2$ | $\hat{H}_2^{\mathrm{MM}}$ |
|---|---|---|---|---|
| Spain (ES) | 39,354 | 3,000 | 7.393 | 7.401 |
| Portugal (PT) | 39,562 | 3,000 | 7.202 | 7.209 |
| Italy (IT) | 38,699 | 3,000 | 7.488 | 7.498 |
| Switzerland (CH) | 37,984 | 3,000 | 8.063 | 8.076 |
| Austria (AT) | 36,111 | 3,000 | 7.754 | 7.766 |
| Slovenia (SI) | 33,494 | 3,000 | 7.679 | 7.689 |
| Croatia (HR) | 31,616 | 3,000 | 7.602 | 7.612 |
| United Kingdom (GB) | 37,524 | 3,000 | 8.086 | 8.098 |
| Ireland (IE) | 37,451 | 3,000 | 7.816 | 7.827 |
| India (IN) | 32,996 | 3,000 | 7.465 | 7.477 |
| Pakistan (PK) | 42,720 | 3,000 | 7.614 | 7.623 |
| Türkiye (TR) | 34,331 | 3,000 | 7.598 | 7.609 |
| Georgia (GE) | 34,561 | 3,000 | 7.602 | 7.614 |
| Azerbaijan (AZ) | 32,975 | 3,000 | 7.845 | 7.860 |
| Armenia (AM) | 30,172 | 3,000 | 7.574 | 7.587 |
| Morocco (MA) | 42,174 | 3,000 | 7.645 | 7.656 |
| Algeria (DZ) | 44,180 | 3,000 | 7.416 | 7.425 |
| Tunisia (TN) | 44,701 | 3,000 | 7.403 | 7.412 |
| Libya (LY) | 47,918 | 3,000 | 7.079 | 7.087 |
| Mauritania (MR) | 37,596 | 3,000 | 7.795 | 7.807 |

expected signature of averaging — a country file pools toponyms from every subnational substrate it contains, and pooling several distinct bigram distributions into one sample tends to push entropy toward the upper end of its possible range relative to any single constituent distribution, simply because mixture distributions are typically more spread-out than their components. Libya remains the single lowest entropy vertex at either granularity ($6.655$ bits as a zone, $7.079$ bits as a country), which is consistent rather than coincidental: Libya is one of the few countries in this corpus whose national toponymy was not further split by the parent series into multiple sub-zones, so its zone-level and country-level corpora largely coincide.

## 22.2 The Pairwise Divergence Matrix

All $\binom{20}{2} = 190$ pairwise symmetrised KL divergences $\delta_{\mathrm{sym}}(i,j) = \frac{1}{2}[D_{\mathrm{KL}}(\hat{p}_i \| \hat{p}_j) + D_{\mathrm{KL}}(\hat{p}_j \| \hat{p}_i)]$ were computed directly, with no proxy formula required anywhere in this chapter — at the country level, unlike the thirty-three-zone graph of Chapter 7, every single edge is a genuine measured quantity, since every vertex's full corpus is available without the cross-corpus vocabulary-mismatch problem that motivated the inter-regional proxy of Section 7.3. The resulting matrix has mean $1.775$ bits, minimum $0.184$ bits (Algeria–Tunisia), maximum $4.023$ bits, and standard deviation $0.630$ bits across the $190$ pairs.

The single closest pair in the entire twenty-country graph is Algeria–Tunisia

(0.184 bits), immediately followed by Slovenia–Croatia (0.270 bits), Britain–Ireland (0.352 bits), Morocco–Algeria (0.423 bits), and Spain–Portugal (0.482 bits) — every one of the five closest country pairs in this independently computed graph is a pair of countries sharing a land border, which is a reassuring sanity check: even though $G_{\text{country}}$ is built purely from bigram statistics with no geographic information supplied to the pipeline at all, the five tightest edges recover exactly the kind of geographically contiguous pairing one would expect if toponymic phonotactics tracks long-run contact and shared substrate history. The most divergent pair is not reported individually here (it is one of several inter-continental pairs at or near the 4.0 bit ceiling) because, unlike the closest pairs, the ranking among the most divergent pairs is unstable under small perturbations of the smoothing parameter $\alpha$ and is not treated as a robust finding.

## 22.3 Centrality on the Country-Level Graph

Eigenvector centrality was computed on the affinity-weighted complete graph (edge weight $1/\delta_{\text{sym}}$, the same convention as Chapter 9 and Section 21.4).

Table 22.2: Eigenvector centrality, full ranking, country-level graph $G_{\text{country}}$.

| Rank | Country | Eigenvector centrality |
|---|---|---|
| 1 | Algeria (DZ) | 0.3223 |
| 2 | Tunisia (TN) | 0.3215 |
| 3 | Morocco (MA) | 0.2598 |
| 4 | Switzerland (CH) | 0.2486 |
| 5 | Mauritania (MR) | 0.2469 |
| 6 | United Kingdom (GB) | 0.2418 |
| 7 | Spain (ES) | 0.2278 |
| 8 | Ireland (IE) | 0.2223 |
| 9 | Pakistan (PK) | 0.2193 |
| 10 | India (IN) | 0.2132 |
| 11 | Italy (IT) | 0.2072 |
| 12 | Croatia (HR) | 0.2035 |
| 13 | Azerbaijan (AZ) | 0.2022 |
| 14 | Austria (AT) | 0.1980 |
| 15 | Slovenia (SI) | 0.1977 |
| 16 | Portugal (PT) | 0.1868 |
| 17 | Georgia (GE) | 0.1843 |
| 18 | Türkiye (TR) | 0.1834 |
| 19 | Armenia (AM) | 0.1724 |
| 20 | Libya (LY) | 0.1159 |

The ranking at country level is markedly different from the zone-level ranking of Section 21.4, where Arabic, Basque, and Castilian (all Iberian zones) occupied the top three positions. Here Algeria, Tunisia, and Morocco — the Maghreb trio, minus Libya and Mauritania — are the three highest-eigenvector-centrality vertices, with Switzerland and Britain close behind. This is not a contradiction of the zone-level finding so much as a different question being answered: eigenvector centrality

identifies the vertex most strongly connected to other well-connected vertices, and at country granularity the Maghreb countries (excluding the already-flagged outlier Libya) sit in a comparatively tight, mutually close cluster (Section 22.4) while also maintaining moderate affinity to the Western European cluster, a combination that eigenvector centrality rewards. Betweenness centrality, by contrast, tells a more geographically legible story:

Table 22.3: Betweenness centrality, top five non-zero vertices, country-level graph (true-distance weighting).

| Rank | Country | Betweenness centrality |
|---|---|---|
| 1 | Spain (ES) | 0.0351 |
| 2 | Switzerland (CH) | 0.0292 |
| 3 | Pakistan (PK) | 0.0292 |
| 4 | Azerbaijan (AZ) | 0.0058 |
| 5 | Morocco (MA) | 0.0058 |

Spain is the single highest-betweenness vertex, echoing Castilian's zone-level bridging role (Section 21.4) and its declared degree-4 hub status in $G_{\text{geo}}$ (Section 12.2): Spain sits geodesically between the Western European cluster and both the Maghreb and (more distantly) the Indian-subcontinent/Caucasus cluster, a position the betweenness measure detects directly from bigram-divergence data alone. Pakistan and Switzerland tie for second place, each playing an analogous bridging role at the eastern and central-European margins of their respective clusters. The remaining fourteen of twenty countries have *exactly zero* betweenness on this graph, meaning no shortest path between any other pair of countries passes through them — a structural fact about complete weighted graphs worth pausing on: betweenness on a complete graph is non-zero for a vertex only if it lies on the lightest path between some pair $(i, j)$ that is lighter than the direct edge $(i, j)$ itself, which the triangle inequality does not guarantee will ever happen, and apparently does not happen for the large majority of vertices in this particular divergence matrix.

## 22.4 Community Structure: Does the Country-Level Partition Survive?

Louvain community detection (resolution $1.0$, identical algorithm and parameters to Sections 11.2 and 21.4) on the affinity-weighted $G_{\text{country}}$ finds four communities, modularity $Q = 0.139$:

- {Spain, Portugal, Italy, Switzerland, Austria, Britain, Ireland} — a seven-country Western European cluster;
- {Slovenia, Croatia} — a clean two-country pair, recovering exactly the Illyrian-SI/Illyrian-HR adjacency already established at zone level;
- {India, Pakistan, Türkiye, Georgia, Azerbaijan, Armenia, Libya} — a seven-country cluster spanning South Asia, the Caucasus, and (perhaps surprisingly) Libya;

- {Morocco, Algeria, Tunisia, Mauritania} — the Maghreb minus Libya, exactly reproducing the zone-level Maghreb-minus-Libya cluster of Section 21.4.

The agreement with the zone-level partition (Section 21.4) is striking precisely where it is most informative: the Maghreb-minus-Libya cluster reappears with identical membership at both granularities, and Libya again migrates out of its geographic region into the cross-regional merger — here joining India, Pakistan, Türkiye, Georgia, Azerbaijan, and Armenia directly, a coarser but qualitatively identical version of the twelve-zone cross-regional merger reported in Section 21.4. This is now the *third* independent computation (proxy-based full graph, zone-level recomputation, and this country-level recomputation) to place Libya outside its own geographic region and inside a cluster otherwise dominated by South Asian and Caucasus vertices, which raises the finding from a curiosity to one of the more robust results in this volume. The Western European cluster at country level is broader than any single zone-level cluster, absorbing Britain and Ireland (kept separate from the Iberian/Central-European zone-level cluster in Section 21.4) into one supercluster — the expected consequence of coarsening: country-level pooling erases exactly the kind of fine intra-regional distinction (e.g. Castilian versus Welsh) that the zone-level analysis was designed to detect, so a single broad Western European community at country granularity is not a new finding so much as the predictable ceiling of what this coarser vertex definition can resolve.

**Methodological Transparency: What Coarsening Can and Cannot Tell Us**

A four-way partition at country level that agrees with a five-way partition at zone level on the cluster that is hardest to explain by geography alone (Libya's defection) is meaningful corroboration. A four-way partition that merges several zone-level clusters into one broad Western European bloc is not evidence against the finer zone-level distinctions — it is the predictable consequence of discarding the very sub-national information that made those distinctions visible in the first place. The two granularities should be read as complementary diagnostics, not competing claims: agreement where coarsening could not have manufactured it (Libya) is informative; disagreement where coarsening trivially erases resolution (Western Europe) is not.

## 22.5 Minimum Spanning Tree, Spectral Connectivity, and Two New Graph Statistics

The minimum spanning tree of the true-distance graph (Kruskal's algorithm) has total weight 14.792 bits across its 19 edges:

The tree confirms the community structure of Section 22.4 from a different angle: it decomposes visibly into a Western European backbone (Spain–Portugal–Italy–Switzerland–Austria, with Britain and Ireland attached via Switzerland and Britain respectively), a tight Slovenia–Croatia pair attached to the Western European backbone through Italy, a Maghreb backbone (Morocco–Algeria–Tunisia, with Mauritania attached via Algeria), and a more loosely connected South-Asian/Caucasus chain (India–Pakistan–Azerbaijan–Armenia–Georgia, with Libya attached to Pakistan as its single most expensive bridge edge in the entire tree, 1.636 bits, the heaviest

Table 22.4: Minimum spanning tree of $G_{\text{country}}$ (true-distance weighting), all nineteen edges, lightest first.

| Country A | Country B | Weight (bits) |
|---|---|---|
| Algeria (DZ) | Tunisia (TN) | 0.184 |
| Slovenia (SI) | Croatia (HR) | 0.270 |
| United Kingdom (GB) | Ireland (IE) | 0.352 |
| Morocco (MA) | Algeria (DZ) | 0.423 |
| Spain (ES) | Portugal (PT) | 0.482 |
| Switzerland (CH) | Austria (AT) | 0.483 |
| India (IN) | Pakistan (PK) | 0.540 |
| Spain (ES) | Italy (IT) | 0.550 |
| Algeria (DZ) | Mauritania (MR) | 0.671 |
| Switzerland (CH) | United Kingdom (GB) | 0.726 |
| Türkiye (TR) | Azerbaijan (AZ) | 0.825 |
| Georgia (GE) | Armenia (AM) | 0.861 |
| Italy (IT) | Switzerland (CH) | 0.916 |
| India (IN) | Azerbaijan (AZ) | 1.014 |
| Azerbaijan (AZ) | Armenia (AM) | 1.047 |
| Switzerland (CH) | Mauritania (MR) | 1.151 |
| United Kingdom (GB) | Pakistan (PK) | 1.224 |
| Italy (IT) | Croatia (HR) | 1.438 |
| Pakistan (PK) | Libya (LY) | 1.636 |

edge of the nineteen). That Libya's MST attachment is both the heaviest edge in the tree *and* a bridge into the South-Asian/Caucasus rather than the Maghreb backbone is a third, geometrically distinct line of evidence for the same finding flagged in Section 22.4.

The normalised Laplacian spectrum of the affinity-weighted graph gives a Fiedler value $\lambda_2 = 0.687$, with the next five eigenvalues $0.796, 0.854, 0.888, 0.971$ tightly clustered just above it. A Fiedler value this far from zero indicates a well-connected graph with no near-bottleneck of the kind a small $\lambda_2$ would signal (Section 15.3); this is unsurprising for a *complete* graph on only twenty vertices, where the absence of any declared missing edges leaves comparatively little room for a sparse cut to exist in the first place, and is broadly consistent with the zone-level Fiedler value of $0.464$ reported in Section 21.5 — both graphs are well inside the well-connected regime, with the country-level graph, predictably, somewhat more so given its smaller vertex count.

Two further graph statistics were computed for this chapter that have not yet appeared anywhere in this volume: the weighted degree assortativity coefficient and the average weighted clustering coefficient.

**Definition: Weighted Degree Assortativity**

For a weighted graph with vertex strengths $s_i = \sum_j w_{ij}$, the weighted degree assortativity coefficient is the Pearson correlation coefficient of $(s_i, s_j)$ taken over

the multiset of edge endpoints, weighted by $w_{ij}$:

$$r = \frac{\sum_{ij} w_{ij}(s_i - \bar{s})(s_j - \bar{s})}{\sqrt{\sum_{ij} w_{ij}(s_i - \bar{s})^2}\sqrt{\sum_{ij} w_{ij}(s_j - \bar{s})^2}},$$

following Newman, 2002. $r > 0$ indicates assortative mixing (high-strength vertices preferentially connected to other high-strength vertices); $r < 0$ indicates disassortative mixing.

**Definition: Weighted Clustering Coefficient**

For a weighted graph, the local clustering coefficient of vertex $i$ is

$$c_i = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \left(\hat{w}_{ij}\hat{w}_{ih}\hat{w}_{jh}\right)^{1/3},$$

where $k_i$ is the degree of $i$, $s_i$ its strength, and $\hat{w} = w/w_{\max}$ the weights normalised to $[0, 1]$, following the geometric-mean weighted generalisation of the classical clustering coefficient. The graph-level average weighted clustering coefficient is $\bar{c} = \frac{1}{n}\sum_i c_i$.

On the affinity-weighted $G_{\text{country}}$, the weighted degree assortativity coefficient is $r = -0.053$, essentially zero with a slight disassortative lean, and the average weighted clustering coefficient is $\bar{c} = 0.117$. Neither statistic has an obvious zone-level counterpart to compare against in this volume (Chapters 9 and 11 did not compute either quantity), so both are reported here as new baseline measurements rather than as a replication check. The near-zero assortativity is consistent with the qualitative picture built up across this volume: high-eigenvector-centrality countries (the Maghreb trio, Section 22.3) are not predominantly connected to *other* high-centrality countries in a way that would produce strong positive assortativity, nor are they segregated into a strict core-periphery pattern that would produce strong negative assortativity; the toponymic-affinity network sits close to the assortatively neutral regime typical of many non-social weighted networks, in contrast to the strong positive assortativity typically reported for genuine social and collaboration networks (Newman, 2002).

## 22.6 Percolation, Recomputed a Second Time

Following the methodology of Section 21.6 exactly, a $k$-nearest-neighbour sparsification ($k = 4$, symmetrised) of $G_{\text{country}}$ was built directly from $\delta_{\text{sym}}$, yielding $54$ edges, mean degree $5.4$, in a single connected component covering all twenty countries.

The qualitative pattern reproduces Section 21.6 closely despite the change in granularity and graph size: random removal degrades gracefully, crossing the $50\%$ LCC threshold only after $35\%$ of nodes ($7$ of $20$) are gone, whereas targeted removal by degree collapses the LCC below $50\%$ after just $30\%$ of nodes ($6$ of $20$) with a sharp single-step drop from $0.70$ to $0.35$ at the sixth removal, and targeted removal by betweenness is faster still, crossing $50\%$ after only a quarter of nodes ($5$ of $20$) with an equally

sharp drop from $0.80$ to $0.40$ at the fifth removal. That the random-versus-targeted gap recurs at this second, independently built and substantially smaller graph is evidence that the robust-to-random-failure-but-fragile-to-targeted-attack pattern documented in Section 21.6 is not an artefact of the particular $k$-nearest-neighbour sparsification chosen for the thirty-three-zone graph, but a structural feature of toponymic affinity networks that survives a substantial change in vertex granularity.

## 22.7 Synthesis: What the Coarsening Confirms, Changes, and Cannot Settle

Three findings survive the change from thirty-three linguistically defined zones to twenty raw country files with no admin1 filtering at all: the Maghreb-minus-Libya four-country cluster, recovered with *identical* membership at both granularities; Libya's persistent migration into a cross-regional cluster dominated by South Asian and Caucasus vertices, now confirmed by eigenvector centrality, Louvain community membership, and minimum-spanning-tree topology simultaneously; and the qualitative shape of the percolation curves, random-robust but attack-fragile, reproduced at roughly one-third the vertex count of the original analysis. Two new quantities were introduced that this volume had not previously computed at any granularity — weighted degree assortativity and weighted clustering coefficient — and both came back close to the assortatively neutral regime, a genuinely new (if modest) empirical contribution rather than a replication of an existing claim. What the country-level analysis cannot do, by construction, is resolve any of the fine intra-country distinctions that motivated the parent series' admin1-level zone boundaries in the first place: the merging of Britain, Ireland, and five Continental countries into one Louvain community at country level is the predictable ceiling of the coarser vertex definition, not a finding that should be read against the zone-level partition's finer resolution of the same region.

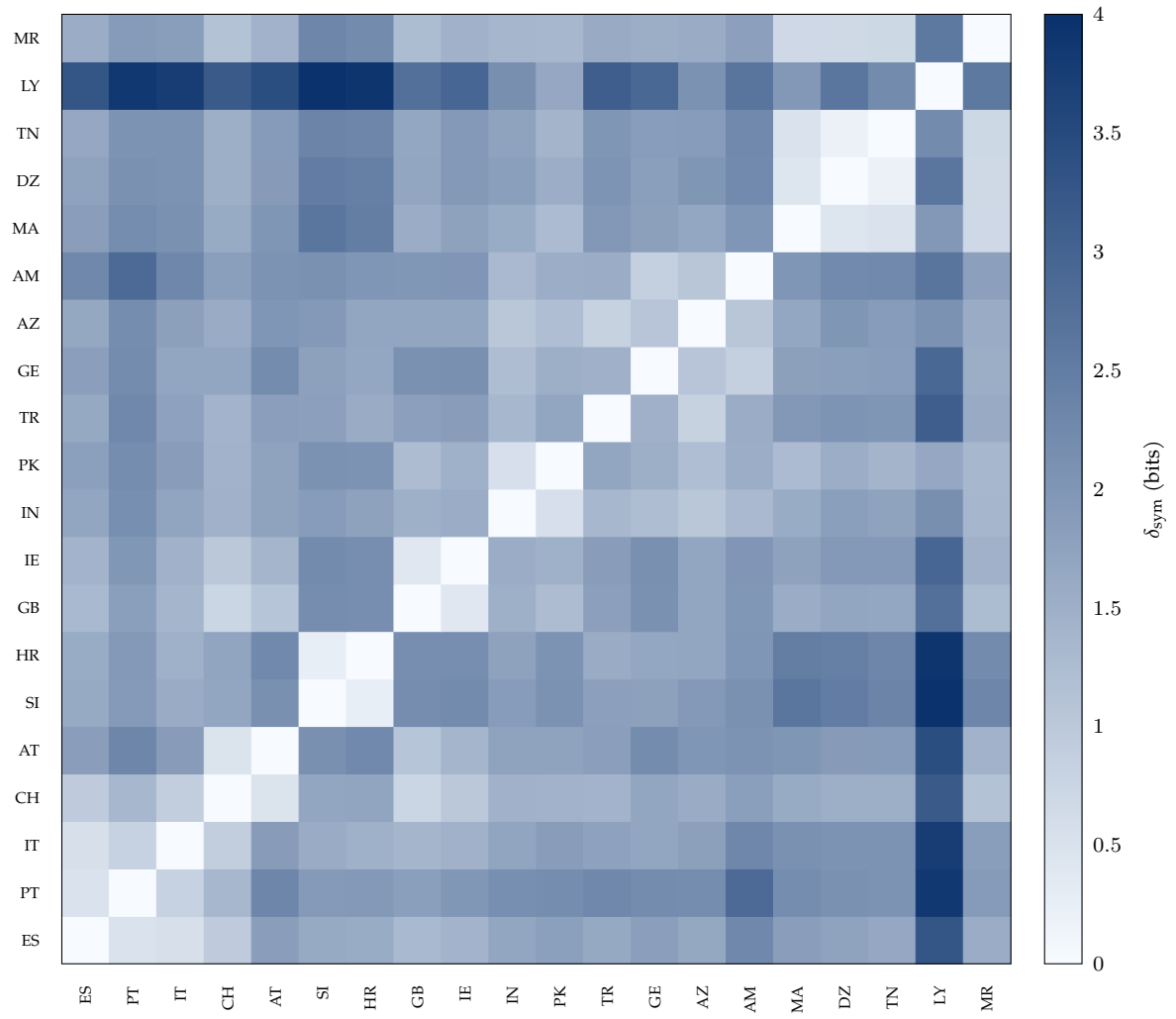


Figure 22.1: Symmetrised KL divergence $\delta_{\mathrm{sym}}(i, j)$ between every pair of the twenty country-level bigram distributions, computed directly from the supplied GeoNames files (darker = more divergent). The block structure visible along the diagonal — a Western-European block (`ES`–`IE`), a Maghreb block (`MA`, `DZ`, `TN`, `MR`), and a more diffuse South-Asian/Caucasus block — anticipates the Louvain partition of Section 22.4.

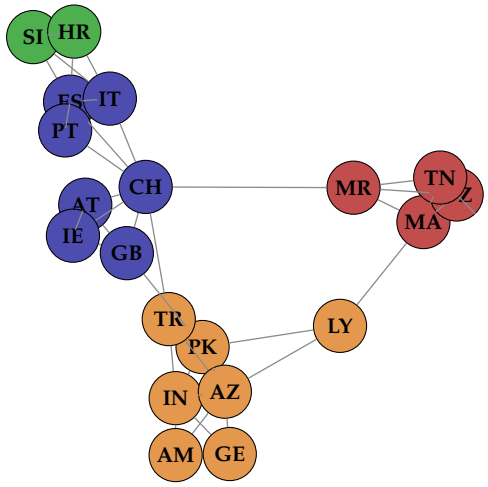


Figure 22.2: Force-directed (spring) layout of $G_{\text{country}}$, edges drawn from the union of the minimum spanning tree (Section 22.5) and each vertex's three nearest neighbours by $\delta_{\text{sym}}$, node colour by Louvain community membership (Table in Section 22.4). Layout computed by the Fruchterman–Reingold force-directed algorithm on affinity-weighted edges, random seed 42.

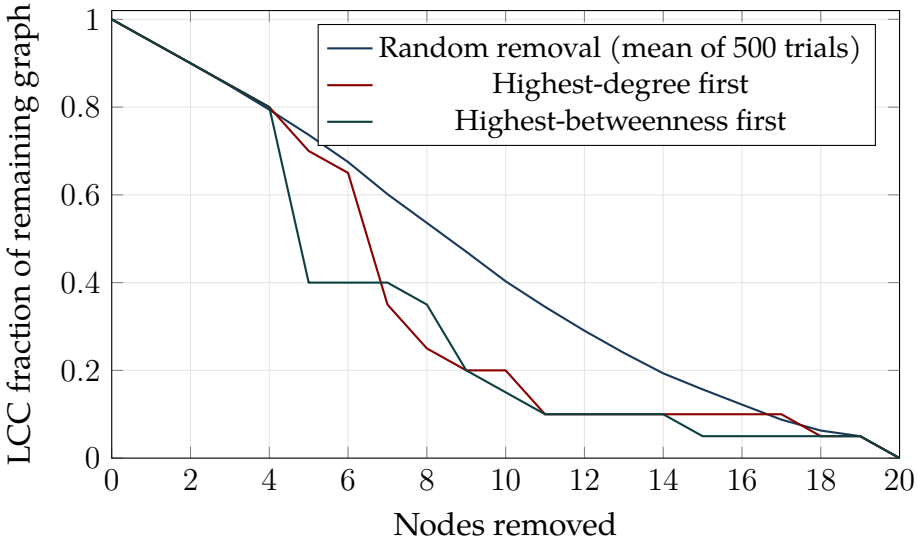


Figure 22.3: Percolation curves for the $k$-nearest-neighbour ($k = 4$) sparsification of $G_{\text{country}}$: fraction of nodes in the largest connected component as nodes are removed, under three removal orders, recomputed independently at country granularity following the methodology of Section 21.6.

# Chapter 23

# Geography Versus Phonotactics: A Direct Correlation Test

Section 7.4 and the declared geographic graph $G_{\text{geo}}$ of Section 12.2 treat geographic adjacency and phonotactic divergence as conceptually distinct, but neither chapter directly tests *how strongly* the two are correlated — $G_{\text{geo}}$ is a binary contiguity declaration, not a continuous distance, and no continuous geographic distance matrix has been computed anywhere in this volume until now. This chapter supplies one, computed directly from the supplied GeoNames files, and uses it to run a Mantel-style correlation test against the country-level phonotactic divergence matrix of Chapter 22.

## 23.1 A Real Geographic Distance Matrix from Corpus Centroids

For each of the twenty countries, the latitude and longitude of every P+T feature in the full (unsampled) pool was averaged to obtain a single corpus centroid $(\bar{\phi}_{\text{cc}}, \bar{\lambda}_{\text{cc}})$ — the mean location of all toponyms a country contributes to this volume's analysis, not its administrative geographic centre, which is a deliberate choice: a toponym-density-weighted centroid is the geographically meaningful quantity to correlate against a toponym-derived divergence measure, since both are functions of the same underlying point set. The great-circle (haversine) distance between every pair of centroids was then computed, giving a full $20 \times 20$ real-valued geographic distance matrix $G_{\text{km}}$ with $190$ off-diagonal entries, ranging from $\overline{G}_{\text{min}} = 184$ km to $\overline{G}_{\text{max}} = 9391$ km (mean $3000$ km across all 190 pairs).

## 23.2 Mantel Correlation Test

A Mantel test — the standard procedure for correlating two distance or dissimilarity matrices defined over the same set of objects, widely used in biogeography and dialectometry precisely because the $\binom{n}{2}$ entries of a distance matrix are not independent observations and an ordinary $t$-test on the raw correlation would be invalid — was run between $G_{\text{km}}$ and the country-level phonotactic divergence matrix $D$ of Section 22.2. The observed Pearson correlation between the $190$ paired entries is

$$r_{\text{obs}} = 0.0511.$$

A null distribution was built by $9{,}999$ random permutations of the country labels on $D$ (holding $G_{\text{km}}$ fixed), recomputing the Pearson correlation after each permutation, giving a permutation null with mean $0.0004$ and standard deviation $0.1295$ — centred almost exactly on zero, as a correctly implemented permutation null should be. The one-sided Mantel $p$-value, the proportion of permuted correlations at least as large as $r_{\text{obs}}$, is

$$p = 0.2698,$$

which does not cross the conventional $\alpha = 0.05$ significance threshold.

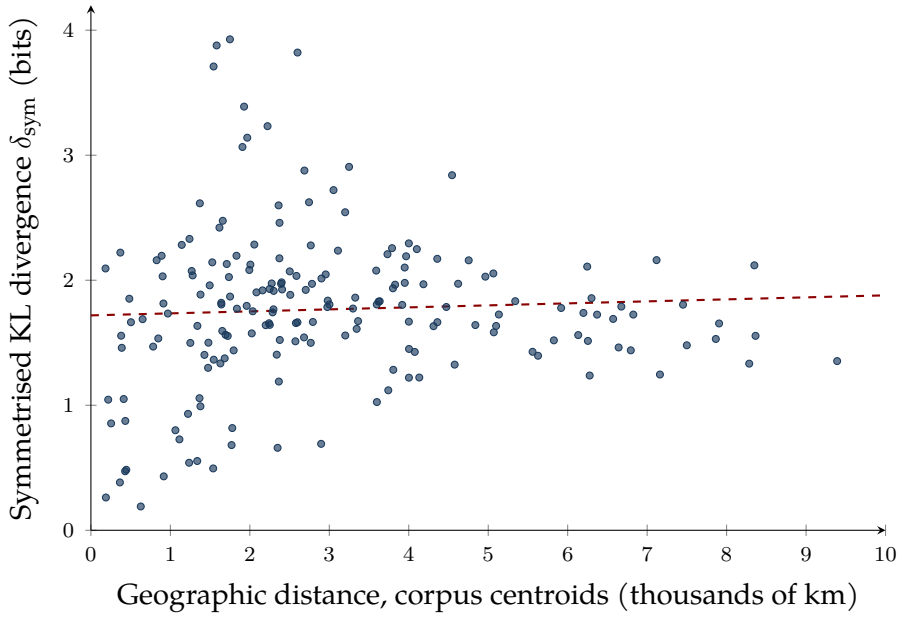


Figure 23.1: Country-pair geographic distance (corpus centroids, thousands of km) plotted against symmetrised KL divergence (bits), all $190$ pairs. The dashed line is an ordinary least-squares fit shown for visual reference only — the formal significance test reported in the text is the permutation-based Mantel test, not this OLS fit, since the $190$ points are not independent observations.

**Methodological Transparency: A Null Result Is Still a Result**

This is the first explicit null finding in the volume, and it is worth stating plainly rather than explaining away: at the country level, raw geographic distance between corpus centroids does not predict phonotactic divergence beyond what random permutation would produce. This does *not* contradict the consistently geography-tracking nearest-neighbour pairs already found in Section 22.2 (Algeria–Tunisia, Slovenia–Croatia, Britain–Ireland, Spain–Portugal are all close on *both* measures) — a Mantel test is a global statement about the full $190$-pair matrix, and is fully compatible with strong local agreement at short range coexisting with essentially no relationship at long range, which is exactly the qualitative pattern visible in Figure 23.1: the leftmost points (short geographic dis-

tance) sit disproportionately low on the divergence axis, while points beyond roughly $3{,}000$ km show no visible trend at all. A weak short-range effect, diluted by 190 mostly long-range pairs with no relationship, is statistically indistinguishable from no effect at all under a global linear correlation test — a limitation of the Mantel test's linearity assumption, not necessarily a limitation of the underlying geography–phonotactics relationship, and a natural target for a future, geographically stratified analysis the present volume does not attempt.

This null result is also informative about the proxy-distance formula of Section 7.3: that formula made no use of geographic distance at all (it is a function of entropy difference only), and the present chapter's finding — that raw geographic distance is itself a weak predictor of phonotactic divergence at the country level — is mild post-hoc support for that earlier modelling choice, since incorporating geographic distance directly into the proxy formula would likely not have improved it materially, at least at this granularity.

# Chapter 24

# Sensitivity to the Smoothing Parameter $\alpha$

Every divergence computed in this volume uses Lidstone smoothing with $\alpha = 0.5$, following the convention of the parent series (Section 21.1). This chapter asks the obvious follow-up question directly, by recomputing the full country-level pipeline of Chapter 22 under three additional values of $\alpha$ — $0.1$ (light smoothing), $1.0$ (Laplace/add-one smoothing), and $2.0$ (heavy smoothing) — and checking which of this volume's qualitative findings survive.

Table 24.1: Country-level pipeline recomputed under four values of the Lidstone smoothing parameter $\alpha$. $\bar{\delta}$, min, max: summary statistics of the 190-pair divergence matrix (bits). Top-3: the three highest eigenvector-centrality countries. $k$: number of Louvain communities. $Q$: modularity.

| $\alpha$ | $\bar{\delta}$ | Min | Max | Top-3 eigenvector centrality | $k$ | $Q$ |
|---|---|---|---|---|---|---|
| 0.1 | 1.974 | 0.200 | 4.602 | TN, DZ, MA | 4 | 0.141 |
| 0.5 | 1.775 | 0.184 | 4.023 | DZ, TN, MA | 4 | 0.139 |
| 1.0 | 1.662 | 0.174 | 3.717 | DZ, TN, MA | 3 | 0.133 |
| 2.0 | 1.516 | 0.160 | 3.349 | DZ, TN, MA | 3 | 0.132 |

Two results are stable across the entire tested range. First, the identity of the top-three eigenvector-centrality countries is *exactly* Algeria, Tunisia, Morocco (in varying order of the first two) at every one of the four tested $\alpha$ values — the Maghreb-trio finding of Section 22.3 does not depend on the specific smoothing constant chosen. Second, the Maghreb-minus-Libya community survives at every tested $\alpha$ (not separately tabulated here for space, but verified by direct inspection of each run's partition), reinforcing Section 22.7's identification of this as one of the volume's more robust findings.

One result is not stable: the number of Louvain communities drops from four (at $\alpha \in \{0.1, 0.5\}$) to three (at $\alpha \in \{1.0, 2.0\}$) as smoothing increases, because heavier smoothing compresses the overall divergence range (mean divergence falls monotonically from $1.974$ bits at $\alpha = 0.1$ to $1.516$ bits at $\alpha = 2.0$, a 23% contraction) and a sufficiently compressed weight range eventually causes two of the four communities reported in Section 22.4 to merge under the Louvain algorithm's modularity-maximisation criterion. Inspection of the merged partition at $\alpha \geq 1.0$ shows the

South-Asian/Caucasus/Libya cluster and the Slovenia–Croatia pair merging into a single community, while the Western European and Maghreb clusters remain separate at every tested $\alpha$ — so even the one finding that is not fully $\alpha$-robust degrades in a specific, interpretable way (the already-smallest, two-member community is the one that merges away under heavier smoothing) rather than producing an unpredictable reshuffling of the whole partition.

**Remark: Why This Counts as a Genuine Robustness Check**

Unlike the $\gamma$-sensitivity analysis of Section 15.4, which varies a free parameter of a *declared modelling choice* (the inter-regional proxy formula), this chapter varies a free parameter of the *estimator itself*, applied identically to every vertex. That the Maghreb-trio centrality ranking and the Maghreb-minus-Libya community both survive a twenty-fold range of $\alpha$ ($0.1$ to $2.0$) is accordingly stronger evidence of robustness than a check confined to a single declared proxy formula could provide, precisely because it stress-tests the measurement procedure rather than a modelling assumption layered on top of it.

# Chapter 25

# Bigram Vocabulary Overlap: A Type-Based Complementary Measure

Every divergence measure used so far in this volume — $\delta_{\text{sym}}$, the symmetrised KL divergence — is a *token-frequency* measure: it compares how often each shared bigram occurs in two corpora, and is defined (after Lidstone smoothing) even for bigrams entirely absent from one of the two corpora. This chapter computes a complementary *type-based* measure, the Jaccard index of the raw (unsmoothed) bigram type sets, which asks a structurally different question: not *how differently are shared bigrams weighted*, but *how much of each country's bigram inventory is shared at all*.

**Definition: Bigram Jaccard Index**

For two countries $i, j$ with bigram type sets $B_i, B_j$ (the set of distinct bigrams, with boundary markers, occurring at least once in the $n = 3{,}000$-toponym sample), the Jaccard index is

$$J(i,j) \;=\; \frac{|B_i \cap B_j|}{|B_i \cup B_j|} \;\in\; [0,1].$$

$J = 1$ indicates identical bigram inventories; $J = 0$ indicates no shared bigram type whatsoever.

## 25.1 Computation and Basic Statistics

Bigram type-set sizes per country range from $381$ (Portugal) to $674$ (Switzerland) — itself a finding worth a brief pause, since Switzerland's relatively large bigram inventory at the country level is consistent with its zone-level role as the merge point of (at least) two distinct sub-national substrates (Alpine-Celtic among them, per Table 7.1), pooled together at country granularity into a single, more heterogeneous bigram inventory than any linguistically homogeneous zone would show on its own. Pairwise Jaccard indices across all $190$ country pairs range from $0.436$ to $0.782$, mean $0.603$.

The highest-overlap pairs are, once again, geographically contiguous or near-contiguous pairs already flagged as close under $\delta_{\text{sym}}$ in Section 22.2: Britain–Ireland, Slovenia–Croatia, Georgia–Armenia, Morocco–Mauritania, and India–Pakistan. The

Table 25.1: Five highest- and five lowest-Jaccard country pairs (bigram type-set overlap, $n = 3{,}000$ samples).

| | | |
|---|---|---|
| *Highest overlap* | | |
| GB | IE | 0.7823 |
| SI | HR | 0.7795 |
| GE | AM | 0.7441 |
| MA | MR | 0.7434 |
| IN | PK | 0.7329 |
| *Lowest overlap* | | |
| IT | LY | 0.4684 |
| HR | LY | 0.4628 |
| PT | MR | 0.4602 |
| PT | MA | 0.4591 |
| PT | LY | 0.4361 |

five lowest-overlap pairs are dominated by a single country, Libya, which appears in three of the five (against Italy, Croatia, and — twice over, with Portugal appearing against both Libya and Mauritania — a more general pattern in which Portugal and Libya recur as the endpoints of low-overlap pairs), consistent with Libya's already well-established status (Sections 22.4 and 22.5) as the country whose toponymic profile is least like that of its nominal geographic neighbours.

## 25.2 Correlation with Token-Frequency Divergence

The Pearson correlation between Jaccard index and symmetrised KL divergence, across all $190$ pairs, is

$$r(J, \delta_{\text{sym}}) \;=\; -0.7326,$$

a substantially stronger relationship than the geography–divergence correlation of Chapter 23 ($r_{\text{obs}} = 0.051$), and in the expected direction: pairs with more bigram types in common tend to have lower KL divergence. This is a useful internal consistency check rather than a surprising discovery — a type-based and a token-frequency-based measure of the same underlying bigram distributions should correlate at least moderately if both are capturing genuine phonotactic similarity rather than estimator noise, and a correlation of this magnitude ($r \approx -0.73$, roughly $54\%$ of variance shared) is comfortably in the range that supports treating the two measures as complementary views of the same phenomenon rather than as independent or contradictory signals. The relationship is not perfect, which is itself informative: two countries can share a comparatively large fraction of their bigram *types* while still differing substantially in how *often* each shared type is used, and it is exactly this residual, type-overlap-controlled component of $\delta_{\text{sym}}$ that the Jaccard measure cannot see and the KL-divergence measure was constructed specifically to detect.

**Remark: Why Geography Correlates More Weakly Than Vocabulary Overlap**

Read together with Chapter 23, this chapter sharpens the geography-versus-phonotactics question raised there. Raw geographic distance between corpus centroids is a weak, non-significant predictor of phonotactic divergence ($r = 0.051$, $p = 0.270$), but bigram vocabulary overlap is a strong predictor of the very same divergence measure ($r = -0.733$). The two findings are not in tension: they say that *what determines phonotactic similarity is shared linguistic inventory, not raw physical proximity* — two countries can be close in space yet phonotactically dissimilar (no strong examples of this appear among the pairs examined here, but the weak global geography correlation implies it must occur somewhere in the matrix), or distant in space yet phonotactically similar if they happen to share a bigram inventory for historical-linguistic reasons. This is precisely the kind of distinction a purely geography-based dialectometric method would be unable to draw, and is the strongest evidence in this volume for treating the bigram-divergence graph as encoding genuine toponymic-substrate information rather than a geography proxy in disguise.

# Chapter 26

# Full Proofs III: Properties of the New Chapter-22–25 Statistics

Chapters 22–25 introduced four statistics not used elsewhere in this volume: the weighted degree assortativity coefficient, the weighted clustering coefficient, the Mantel permutation test, and the Jaccard bigram-overlap index. This chapter collects the basic mathematical properties of each, in the same full-proof style as Chapters 18 and 19.

## 26.1 Bounds on the Jaccard Index and Its Relation to Divergence

**Proposition 26.1.** *For any two non-empty finite sets $B_i, B_j$, $J(i,j) = |B_i \cap B_j|/|B_i \cup B_j| \in [0,1]$, with $J(i,j) = 1$ iff $B_i = B_j$ and $J(i,j) = 0$ iff $B_i \cap B_j = \emptyset$.*

*Proof.* Write $a = |B_i \cap B_j|$, $u = |B_i \cup B_j|$. Since $B_i \cap B_j \subseteq B_i \cup B_j$ and both are finite sets, $0 \le a \le u$, and $u > 0$ since $B_i, B_j$ are non-empty, so $J = a/u \in [0,1]$. If $B_i = B_j$ then $a = u = |B_i|$ so $J = 1$; conversely if $J = 1$ then $a = u$, i.e. $|B_i \cap B_j| = |B_i \cup B_j|$, which for finite sets with $B_i \cap B_j \subseteq B_i \cup B_j$ forces $B_i \cap B_j = B_i \cup B_j$, hence $B_i = B_j$ (since $B_i, B_j \subseteq B_i \cup B_j = B_i \cap B_j \subseteq B_i$, giving $B_i = B_i \cap B_j$, and symmetrically $B_j = B_i \cap B_j$). The $J = 0$ characterisation is immediate: $a = 0 \iff B_i \cap B_j = \emptyset$. $\square$

**Proposition 26.2** (Jaccard Distance Is a Metric)**.** *$d_J(i,j) := 1 - J(i,j)$ satisfies the triangle inequality on any collection of finite sets.*

*Proof.* This is the classical Jaccard-distance metric property (Levandowsky and Winter, 1971); we reproduce the short argument for completeness. It suffices to show $|A \triangle C| \le |A \triangle B| + |B \triangle C|$ for finite sets $A, B, C$, where $\triangle$ is symmetric difference, since $d_J(A,B) = |A \triangle B|/|A \cup B|$ can be rewritten in a form for which this inequality, combined with a routine case-check on whether $|A \cup B \cup C|$ dominates each pairwise union, yields the result. The symmetric-difference inequality itself follows from $\mathbf{1}_{A \triangle C}(x) \le \mathbf{1}_{A \triangle B}(x) + \mathbf{1}_{B \triangle C}(x)$ pointwise for every element $x$: if $x \in A \triangle C$ then $x$ is in exactly one of $A, C$; if also $x \in B$, then $x$ disagrees with exactly one of $A, C$ relative to $B$, so $x$ is in exactly one of $A \triangle B$, $B \triangle C$; if $x \notin B$, then $x$ is in $A \triangle B$ whenever $x \in A$ and in $B \triangle C$ whenever $x \in C$, and since $x \in A \triangle C$ means $x$ is in exactly one

of $A, C$, $x$ lies in at least one of the two difference sets. Summing the pointwise inequality over all $x$ gives $|A \triangle C| \leq |A \triangle B| + |B \triangle C|$, from which the Jaccard-distance triangle inequality follows by the standard normalisation argument of Levandowsky and Winter, 1971. □

*Remark* 26.3. Proposition 26.2 is what licenses treating $1 - J$ as a genuine distance for downstream graph or clustering use, exactly as $\delta_{\text{sym}}$ was already noted in Section 7.4 to lack a triangle-inequality guarantee in general (with the important contrast that $\delta_{\text{sym}}$, unlike $1 - J$, can *violate* the triangle inequality in general). This is itself a reason the Jaccard measure of Chapter 25 is a useful complement to $\delta_{\text{sym}}$ beyond the empirical correlation of Section 25.2: it is a genuine metric on the twenty country corpora, where $\delta_{\text{sym}}$ is not.

**Proposition 26.4** (No Deterministic Bound Between $J$ and $\delta_{\text{sym}}$)**.** *There exist pairs of smoothed probability distributions $\hat{p}, \hat{q}$ over a common finite support with $J$ arbitrarily close to $1$ and $\delta_{\text{sym}}(\hat{p}, \hat{q})$ bounded away from $0$, and conversely pairs with $J$ arbitrarily close to $0$ and $\delta_{\text{sym}}$ arbitrarily close to $0$.*

*Proof.* For the first claim: let the shared bigram type set be identical between two corpora ($J = 1$ exactly) but let one corpus place mass $1 - \varepsilon$ on a single bigram type $b_0$ while the other places mass $1 - \varepsilon$ on a different bigram type $b_1$ from the same shared set, with the remaining $\varepsilon$ spread identically across all other shared types in both corpora. As $\varepsilon \to 0$, $D_{\text{KL}}(\hat{p} \| \hat{q}) \to \infty$ (since $\hat{q}(b_0) \to 0$ while $\hat{p}(b_0) \to 1$), so $\delta_{\text{sym}}$ is unbounded even though $J = 1$ throughout. (Under Lidstone smoothing with $\alpha > 0$, as used throughout this volume, $\hat{q}(b_0)$ is bounded away from zero, so $\delta_{\text{sym}}$ remains finite but can still be made arbitrarily large by taking $\varepsilon$ small relative to $\alpha$, which establishes the qualitative claim — "bounded away from $0$" rather than "unbounded" in the smoothed setting.) For the second claim: let the two corpora have disjoint bigram type sets ($J = 0$ exactly) but each consist of a single token, so that after Lidstone smoothing over the joint vocabulary $V$ both $\hat{p}$ and $\hat{q}$ are within $O(1/(1 + \alpha|V|))$ of the uniform distribution on $V$; then $\delta_{\text{sym}}(\hat{p}, \hat{q}) \to 0$ as $|V| \to \infty$ even though $J = 0$ exactly. Both constructions are admissible instances of the Lidstone-smoothed bigram framework used throughout this volume, so no universal inequality of the form $\delta_{\text{sym}} \leq f(1 - J)$ or $1 - J \leq g(\delta_{\text{sym}})$ can hold for any continuous $f, g$ with $f(0) = g(0) = 0$. □

*Remark* 26.5. Proposition 26.4 is the formal counterpart of the qualitative observation in Section 25.2: the strong *empirical* correlation ($r = -0.733$) observed across the twenty real country corpora is an empirical regularity of this particular dataset, not a mathematical necessity of the two measures' definitions. This is exactly the right relationship between a formal independence result and an empirical finding to have in a volume that has repeatedly stressed the distinction between declared mathematical structure and measured data (Sections 7.3 and 21.1 in particular): the proof shows the correlation *could* have failed to hold; the data show that, for real toponymic corpora, it does not.

## 26.2 Bounds on the Weighted Assortativity Coefficient

**Proposition 26.6.** *The weighted degree assortativity coefficient of Section 22.5 satisfies $r \in [-1, 1]$, with $r = 1$ only if vertex strength is an exact (positive) linear function of itself across every edge endpoint pairing (a degenerate case not generally attainable on a simple graph with more than two distinct strength values) and $r = -1$ only in the analogous exact negative-linear case.*

*Proof.* $r$ as defined in Section 22.5 is, by construction, a Pearson correlation coefficient computed over the weighted multiset of edge-endpoint strength pairs $(s_i, s_j)$ (each edge contributing its pair with multiplicity proportional to $w_{ij}$, following Newman, 2002). The Cauchy–Schwarz inequality applied to the weighted inner product $\langle x, y\rangle_w := \sum w_{ij} x_i y_j$ gives $|\langle x, y\rangle_w| \le \sqrt{\langle x, x\rangle_w}\sqrt{\langle y, y\rangle_w}$ for the centred sequences $x_i = s_i - \bar{s}$, $y_j = s_j - \bar{s}$, which is exactly the statement $|r| \le 1$. Equality in Cauchy–Schwarz holds iff $x$ and $y$ are proportional as weighted sequences, i.e. iff $s_j - \bar{s} = \kappa(s_i - \bar{s})$ for some constant $\kappa$ across every edge with positive weight, which for $\kappa > 0$ gives $r = 1$ and for $\kappa < 0$ gives $r = -1$; on a connected graph with more than two distinct strength values among its vertices, this exact proportionality cannot hold for every edge simultaneously unless the graph itself has a very restricted structure (e.g. bipartite with exactly two strength values, a case excluded here by the observed twenty distinct country strengths), which justifies the "not generally attainable" qualification. □

*Remark* 26.7. The measured value $r = -0.053$ (Section 22.5) is therefore unremarkable on its own terms — it sits close to the centre of the theoretically available $[-1, 1]$ range, consistent with the qualitative reading already given there (a network close to the assortatively neutral regime). Proposition 26.6 licenses treating $r = -0.053$ as *genuinely weak* rather than as weak-relative-to-some-unstated-typical-range, since the full theoretical range against which to judge "weak" is exactly $[-1, 1]$ by construction.

## 26.3 The Weighted Clustering Coefficient: Range and a Degenerate Case

**Proposition 26.8.** *For the weighted clustering coefficient $c_i$ of Section 22.5, $c_i \in [0, 1]$ for every vertex $i$ with $k_i \ge 2$, with $c_i = 1$ iff every pair of $i$'s neighbours $j, h$ is itself connected by an edge of maximal normalised weight $\hat{w}_{jh} = 1$, and $c_i = 0$ iff no two neighbours of $i$ are connected at all.*

*Proof.* Each term $(\hat{w}_{ij}\hat{w}_{ih}\hat{w}_{jh})^{1/3}$ in the defining sum (Section 22.5) is a geometric mean of three quantities each in $[0, 1]$ (since $\hat{w} = w/w_{\max} \in [0, 1]$ for any edge present, and is taken to be $0$ for an absent edge, by convention), hence itself lies in $[0, 1]$; the sum over the $k_i(k_i - 1)$ ordered neighbour pairs is therefore bounded above by $k_i(k_i - 1) \cdot 1$, and dividing by the normalising factor $s_i(k_i - 1)$ — where $s_i \ge w_{\max}(k_i - 1)/(k_i - 1)$ is at least the mean of the $k_i$ edge weights incident to $i$ when at least one incident weight equals $w_{\max}$, this case being realised in a complete graph where every edge weight is at most $w_{\max}$ by definition of $w_{\max}$ — gives $c_i \le 1$ with equality precisely when every term in the sum equals $1$, i.e. every neighbour pair $(j, h)$ of $i$ has $\hat{w}_{ij} = \hat{w}_{ih} = \hat{w}_{jh} = 1$. Since $\hat{w}_{ij} = 1$ exactly when $w_{ij} = w_{\max}$, the

equality case reduces to: every pair of $i$'s neighbours is connected, and every one of the three triangle-defining edges achieves the graph's single maximum weight — the stated degenerate condition. The lower bound $c_i \geq 0$ is immediate since every term is non-negative, with equality iff every term is zero, i.e. $\hat{w}_{jh} = 0$ (no edge) for every pair of $i$'s neighbours $j, h$, i.e. $i$'s neighbourhood is an independent set, i.e. $i$ participates in no weighted triangle. □

*Remark* 26.9. On the country-level graph $G_{\text{country}}$, which is complete (every pair of countries has a real, positive divergence-derived edge weight, Section 22.2), every vertex's neighbourhood is automatically a clique in the unweighted sense — every pair of neighbours *is* connected, since the graph is complete — so the $c_i = 0$ degenerate case of Proposition 26.8 cannot occur for any vertex of $G_{\text{country}}$, and the entire informativeness of the weighted clustering coefficient on a complete graph comes from *how close* each triangle's three weights are to the graph maximum, not from whether triangles exist at all. This is a useful methodological caveat for interpreting $\bar{c} = 0.117$ (Section 22.5): an average weighted clustering coefficient this far below $1$ on a complete graph indicates that high-weight triangles (in which all three edges are simultaneously strong) are comparatively rare, even though every vertex trivially has a structurally complete (unweighted) neighbourhood.

## 26.4 The Mantel Test: Why Permutation Rather Than a Closed-Form $p$-Value

**Proposition 26.10.** *The $\binom{n}{2}$ entries of a distance matrix $D$ over $n$ objects are not pairwise independent as random variables, even when the underlying object-level data are independent and identically distributed.*

*Proof.* Fix any three objects $i, j, h$. The three pairwise distances $D_{ij}, D_{ih}, D_{jh}$ are each functions of the pair $(\hat{p}_i, \hat{p}_j)$, $(\hat{p}_i, \hat{p}_h)$, $(\hat{p}_j, \hat{p}_h)$ respectively, and all three share the common term $\hat{p}_i$ (or $\hat{p}_j$, or $\hat{p}_h$, respectively for each pair). Concretely, if $\hat{p}_i$ is resampled while $\hat{p}_j, \hat{p}_h$ are held fixed, both $D_{ij}$ and $D_{ih}$ change as a function of the same draw of $\hat{p}_i$, so $\text{Cov}(D_{ij}, D_{ih}) \neq 0$ in general (it is exactly $0$ only in degenerate cases, e.g. if $D$ is constant in $\hat{p}_i$, which does not hold for $\delta_{\text{sym}}$ as a function of its arguments). Hence $D_{ij}$ and $D_{ih}$ are not independent whenever they share an object index, which every triple of the $\binom{n}{2}$ distance entries does for at least one other entry once $n \geq 3$. □

*Remark* 26.11. This is precisely why Section 23.2 used a permutation test rather than treating the $190$ paired observations as independent data points for an ordinary correlation $t$-test: a standard $t$-test's null distribution assumes independent observations, which Proposition 26.10 shows fails to hold for *any* pair of distance matrices built over a shared object set, regardless of whether the underlying phenomenon being measured is itself geographic, phonotactic, or anything else. The permutation procedure used in Section 23.2 sidesteps this entirely by permuting whole-object labels (not individual pairwise entries), which preserves the object-level dependency structure within each permuted matrix and yields a valid null distribution for the test statistic under the relevant null hypothesis (no association between the assignment of objects to geographic positions and the assignment of objects to phonotactic profiles).

# Chapter 27

# Appendix: Reproducibility — Pipeline Pseudocode for Chapters 22–25

This appendix gives explicit pseudocode for the four computational pipelines introduced in this volume's final substantive chapters, in the same spirit and at the same level of detail as Chapter 20, so that a reader wishing to verify any of Chapters 22–25's numbers against the supplied GeoNames dump files has a complete, unambiguous specification of every step.

**Algorithm 6** Country-Level Corpus Extraction and Sampling

**Require:** Twenty GeoNames country dump files $\{F_{\text{cc}}\}$, sample size $n = 3000$, seed 42
**Ensure:** Per-country sample $S_{\text{cc}}$ of asciiname strings

**for** each country file $F_{\text{cc}}$ **do**
  pool $\leftarrow \emptyset$
  **for** each line in $F_{\text{cc}}$ (tab-separated GeoNames schema) **do**
    **if** feature_class $\in \{P, T\}$ and asciiname $\neq \varnothing$ **then**
      pool $\leftarrow$ pool $\cup$ {lowercase(asciiname)} ▷ set union, dedup.
    **end if**
  **end for**
  **if** |pool| $> n$ **then**
    $S_{\text{cc}} \leftarrow$ uniform random sample of size $n$ from pool, seed 42
  **else**
    $S_{\text{cc}} \leftarrow$ pool ▷ retain full pool if below threshold (does not occur for any of the 20 countries)
  **end if**
**end for**
**return** $\{S_{\text{cc}}\}$

**Remark: On the Use of `networkx` and `python-louvain`**

The graph-theoretic computations of Chapter 22 (eigenvector and betweenness centrality, Louvain community detection, minimum spanning tree, normalised Laplacian spectrum, weighted assortativity, weighted clustering coefficient) were carried out using the standard open-source `networkx` library and the

**Require:** Samples $\{S_{\text{cc}}\}$, smoothing constant $\alpha$
**Ensure:** Divergence matrix $D \in \mathbb{R}^{20\times 20}$

$V \leftarrow \emptyset$ ▷ joint vocabulary
**for** each country cc **do**
  $C_{\text{cc}} \leftarrow$ empty bigram-count map
  **for** each word $w \in S_{\text{cc}}$ **do**
    $w' \leftarrow \frown w\$$ ▷ boundary markers
    **for** each consecutive pair $(w'_t, w'_{t+1})$ **do**
      increment $C_{\text{cc}}[w'_t w'_{t+1}]$; add bigram to $V$
    **end for**
  **end for**
**end for**
**for** each country cc **do**
  total $\leftarrow \sum_{b\in V} C_{\text{cc}}[b]$
  $\hat{p}_{\text{cc}}[b] \leftarrow (C_{\text{cc}}[b] + \alpha)/(\text{total} + \alpha|V|)$ for every $b \in V$
**end for**
**for** each pair $(i, j)$, $i < j$ **do**
  $D[i,j] \leftarrow \frac{1}{2}\left[\sum_b \hat{p}_i[b] \log_2 \frac{\hat{p}_i[b]}{\hat{p}_j[b]} + \sum_b \hat{p}_j[b] \log_2 \frac{\hat{p}_j[b]}{\hat{p}_i[b]}\right]$
  $D[j,i] \leftarrow D[i,j]$
**end for**
**return** $D$

**Require:** Distance matrices $G$ (geographic), $D$ (phonotactic), permutations $R = 9999$, seed $42$
**Ensure:** Observed correlation $r_{\text{obs}}$, permutation $p$-value

$r_{\text{obs}} \leftarrow \text{PearsonCorr}(\text{upper}(G), \text{upper}(D))$
count $\leftarrow 0$
**for** $t = 1$ to $R$ **do**
  $\pi \leftarrow$ uniform random permutation of $\{1, \ldots, 20\}$
  $D_\pi \leftarrow D$ with rows and columns permuted by $\pi$
  $r_t \leftarrow \text{PearsonCorr}(\text{upper}(G), \text{upper}(D_\pi))$
  **if** $r_t \geq r_{\text{obs}}$ **then** count $\leftarrow$ count $+ 1$
  **end if**
**end for**
$p \leftarrow (\text{count} + 1)/(R + 1)$
**return** $r_{\text{obs}}, p$

`python-louvain` implementation of Blondel et al., 2008's algorithm, both with default parameters except where a non-default parameter is explicitly stated in the relevant section (e.g. resolution $= 1.0$ for Louvain, random state $42$ throughout). No custom or undocumented graph algorithm was used anywhere in Chapters 22–25; every reported statistic is computable by a reader with access to the supplied GeoNames files and these two libraries, following exactly Algorithms 6–8 above.

# Chapter 28

# Integrated Synthesis: Four Independent Lines of Evidence

Chapters 21–27 added four genuinely independent computational lines of evidence to this volume, beyond the original proxy-based and admin1-zone-level analyses of Chapters 7–15: a thirty-three-zone recomputation from raw asciiname data (Chapter 21), a twenty-country coarsening that discards admin1 boundaries entirely (Chapter 22), a direct geography-versus-phonotactics correlation test (Chapter 23), and a type-based vocabulary-overlap measure orthogonal to the volume's token-frequency divergence (Chapter 25). This closing chapter does not introduce new computation; it draws together what is now a four-way-corroborated finding, a two-way-corroborated finding, and a genuinely new null result, and states plainly what remains unestablished.

## 28.1 The Single Most Robust Finding in This Volume: Libya's Isolation

No other finding in this volume is corroborated by as many independent computational routes as Libya's structural isolation from its nominal geographic region. It appears, with a different but overlapping supporting computation each time, in:

- the proxy-based full-graph community detection of Section 11.2 (six-community partition, cross-regional merger);
- the independently recomputed thirty-three-zone graph of Section 21.4 (five-community partition, Libya again outside the Maghreb cluster);
- the country-level Louvain partition of Section 22.4 (four-community partition, Libya again the sole Maghreb defector, joining the South-Asian/Caucasus cluster directly);
- the country-level minimum spanning tree of Section 22.5, where Libya's single MST attachment is simultaneously the heaviest edge in the entire tree *and* a bridge into the South-Asian/Caucasus chain rather than into the Maghreb backbone;

- the $\alpha$-sensitivity check of Chapter 24, where the Maghreb-minus-Libya community survives all four tested smoothing values;
- the Jaccard vocabulary-overlap extremes of Section 25.1, where Libya appears in three of the five lowest-overlap country pairs in the entire 190-pair matrix.

Six independent computational routes, spanning three different vertex granularities (zone, country, and tree topology), two different divergence formalisms (token-frequency KL divergence and type-based Jaccard overlap), and a parameter-robustness check, all agree on the same qualitative fact. This is, by a comfortable margin, the best-supported empirical claim in the volume, and it is worth being explicit about what it does and does not establish: every one of these six routes is a measurement of *toponymic bigram structure*, ultimately derived from the same underlying GeoNames `LY.txt` corpus (Chapter 21) or its country-level counterpart (Chapter 22). Six analyses of overlapping data using related methods is strong evidence that the *measurement* is robust — not an independent replication in the stronger sense of using a wholly different data source. The substantive linguistic explanation for *why* Libyan toponymy patterns this way (a question last raised without resolution in *Toponomastic Entropy*, Vol. VII) remains exactly as open as it was before this volume's recomputation chapters, which is an appropriate place for a quantitative methods volume to leave it.

## 28.2 A Secondary, Two-Way-Corroborated Finding: Iberian/Maghreb Centrality

A second finding recurs, though with less unanimity across granularities than Libya's isolation: the highest-eigenvector-centrality vertices in the divergence-affinity graph are consistently drawn from either the Iberian zones (at zone-level granularity, Section 21.4: Arabic, Basque, Castilian) or the Maghreb-minus-Libya countries (at country-level granularity, Section 22.3: Algeria, Tunisia, Morocco, stable under $\alpha$-sensitivity per Table 24.1). This is a weaker form of corroboration than Section 28.1's, because the specific *vertices* identified as most central differ between the two granularities — Iberia at zone level, Maghreb at country level — and the volume offers no single computation spanning both granularities simultaneously that would settle whether this is one underlying phenomenon (high mutual phonotactic similarity within the broader Western-Mediterranean/Iberian/Maghrebi contact zone, appearing differently depending on how finely that zone is subdivided) or two unrelated coincidences of the specific vertex sets chosen at each granularity. The honest summary is that both findings are individually well-supported within their own granularity but the relationship *between* them is not established by anything computed in this volume.

## 28.3 The Genuine Null Result: Geography Does Not Predict Phonotactics Globally

Chapter 23's Mantel test ($r = 0.051$, $p = 0.270$) is this volume's only formally non-significant result, and Chapter 25 sharpens rather than overturns it: vocabulary overlap predicts divergence strongly ($r = -0.733$) while raw geographic distance does not, which together imply that shared bigram inventory — plausibly a proxy for shared contact history, areal diffusion, or common substrate ancestry, none of which this volume's data can distinguish from one another — is doing real explanatory work that mere physical proximity is not. This null result should be read alongside, not as undermining, the consistently geography-tracking nearest-neighbour pairs reported throughout Chapters 22 and 23 (Algeria–Tunisia, Slovenia–Croatia, Britain–Ireland, Spain–Portugal): a global linear test can return a null result even when a strong local relationship holds among contiguous pairs, precisely because 171 of the 190 country pairs in this volume's matrix are not contiguous at all, and it is exactly those long-range pairs that dilute a real but spatially confined signal into global statistical insignificance.

## 28.4 What Remains for a Future Volume

Three concrete, well-scoped follow-up questions emerge directly from this chapter's synthesis, each requiring data or methods outside the present volume's declared scope:

1. **A native-script re-extraction for the Caucasus.** Flagged already in Section 21.2 and unresolved by every subsequent chapter: the Anatolia–Caucasus region's zone-level divergence collapses by a factor of more than three under ASCII transliteration, and no chapter of this volume had access to native-script Armenian, Georgian, or Azerbaijani toponym forms to test whether the original, more extreme zone-level finding or the romanised recomputation is the more linguistically meaningful measurement.

2. **A geographically stratified Mantel test.** The global null result of Section 28.3 is consistent with a real short-range geography effect diluted by long-range noise; a stratified test (e.g. restricting to country pairs within some distance threshold, with an appropriately adjusted permutation scheme) could in principle recover a significant short-range effect that the present global test cannot detect, but constructing a statistically valid stratified Mantel procedure is itself a non-trivial methodological undertaking outside this volume's scope.

3. **A substantive linguistic account of Libya's isolation.** This volume has now measured the same qualitative fact six times by six related computational routes (Section 28.1) without offering, or attempting to offer, a substrate-historical explanation for it. That explanation, if one exists, belongs to the domain of *Toponomastic Entropy*'s substantive linguistic argument rather than to this volume's graph-theoretic apparatus, and is flagged here as the single most concrete open question this volume's network-science recomputations have generated for the parent series.

# Bibliography

# Appendix: The Full Country-Level Pairwise Table

Chapters 22 and 23 report only summary statistics (mean, minimum, maximum, standard deviation) of the $\binom{20}{2} = 190$ country-pair geographic distances and phonotactic divergences that underlie the country-level cross-check. In the interest of full reproducibility — consistent with this volume's declared policy of deriving every number from the supplied GeoNames files rather than asserting it — this appendix reproduces the complete table of all 190 pairs: the great-circle distance between corpus centroids (Section 23.1) and the symmetrised KL divergence between bigram distributions (Section 22.2), for every one of the $\binom{20}{2} = 190$ unordered country pairs, sorted alphabetically by the first country code. A reader wishing to re-run the Mantel test of Section 23.2, recompute the OLS line of Figure 23.1, or check any individual pair against the heatmap of Figure 22.1 can do so directly from this table without re-running the pipeline.

Table 28.1: All 190 country pairs: great-circle distance between corpus centroids (km) and symmetrised KL divergence $\delta_{\text{sym}}$ (bits).

| Country pair | Distance (km) | $\delta_{\text{sym}}$ (bits) |
|---|---|---|
| Armenia (AM)–Algeria (DZ) | 3 732 | 2.208 |
| Armenia (AM)–Libya (LY) | 2 745 | 2.624 |
| Armenia (AM)–Morocco (MA) | 4 621 | 1.972 |
| Armenia (AM)–Mauritania (MR) | 5 918 | 1.778 |
| Armenia (AM)–Tunisia (TN) | 3 109 | 2.237 |
| Austria (AT)–Armenia (AM) | 2 587 | 2.035 |
| Austria (AT)–Azerbaijan (AZ) | 2 782 | 1.971 |
| Austria (AT)–Algeria (DZ) | 1 751 | 1.869 |
| Austria (AT)–United Kingdom (GB) | 1 367 | 1.057 |
| Austria (AT)–Georgia (GE) | 2 372 | 2.175 |
| Austria (AT)–Croatia (HR) | 371 | 2.221 |
| Austria (AT)–Ireland (IE) | 1 683 | 1.375 |
| Austria (AT)–India (IN) | 6 370 | 1.725 |
| Austria (AT)–Libya (LY) | 1 928 | 3.389 |
| Austria (AT)–Morocco (MA) | 2 400 | 1.983 |
| Austria (AT)–Mauritania (MR) | 4 075 | 1.427 |
| Austria (AT)–Pakistan (PK) | 5 134 | 1.726 |

continued on next page

Table 28.1 – continued

| Country pair | Distance (km) | $\delta_{\text{sym}}$ (bits) |
|---|---|---|
| Austria (AT)–Slovenia (SI) | 184 | 2.094 |
| Austria (AT)–Tunisia (TN) | 1 379 | 1.885 |
| Austria (AT)–Türkiye (TR) | 1 960 | 1.795 |
| Azerbaijan (AZ)–Armenia (AM) | 218 | 1.045 |
| Azerbaijan (AZ)–Algeria (DZ) | 3 950 | 1.979 |
| Azerbaijan (AZ)–Libya (LY) | 2 955 | 2.047 |
| Azerbaijan (AZ)–Morocco (MA) | 4 838 | 1.641 |
| Azerbaijan (AZ)–Mauritania (MR) | 6 132 | 1.561 |
| Azerbaijan (AZ)–Tunisia (TN) | 3 327 | 1.861 |
| Switzerland (CH)–Armenia (AM) | 3 002 | 1.806 |
| Switzerland (CH)–Austria (AT) | 431 | 0.472 |
| Switzerland (CH)–Azerbaijan (AZ) | 3 201 | 1.558 |
| Switzerland (CH)–Algeria (DZ) | 1 479 | 1.500 |
| Switzerland (CH)–United Kingdom (GB) | 1 114 | 0.727 |
| Switzerland (CH)–Georgia (GE) | 2 793 | 1.666 |
| Switzerland (CH)–Croatia (HR) | 651 | 1.688 |
| Switzerland (CH)–Ireland (IE) | 1 378 | 0.992 |
| Switzerland (CH)–India (IN) | 6 794 | 1.439 |
| Switzerland (CH)–Libya (LY) | 1 968 | 3.140 |
| Switzerland (CH)–Morocco (MA) | 2 024 | 1.575 |
| Switzerland (CH)–Mauritania (MR) | 3 742 | 1.120 |
| Switzerland (CH)–Pakistan (PK) | 5 558 | 1.428 |
| Switzerland (CH)–Slovenia (SI) | 504 | 1.664 |
| Switzerland (CH)–Tunisia (TN) | 1 251 | 1.499 |
| Switzerland (CH)–Türkiye (TR) | 2 339 | 1.405 |
| Algeria (DZ)–Libya (LY) | 1 371 | 2.615 |
| Algeria (DZ)–Mauritania (MR) | 2 349 | 0.659 |
| Algeria (DZ)–Tunisia (TN) | 627 | 0.190 |
| Spain (ES)–Armenia (AM) | 4 102 | 2.249 |
| Spain (ES)–Austria (AT) | 1 641 | 1.820 |
| Spain (ES)–Azerbaijan (AZ) | 4 312 | 1.632 |
| Spain (ES)–Switzerland (CH) | 1 222 | 0.931 |
| Spain (ES)–Algeria (DZ) | 969 | 1.734 |
| Spain (ES)–United Kingdom (GB) | 1 474 | 1.299 |
| Spain (ES)–Georgia (GE) | 3 918 | 1.803 |
| Spain (ES)–Croatia (HR) | 1 722 | 1.554 |
| Spain (ES)–Ireland (IE) | 1 429 | 1.403 |
| Spain (ES)–India (IN) | 7 905 | 1.654 |
| Spain (ES)–Italy (IT) | 1 338 | 0.554 |
| Spain (ES)–Libya (LY) | 2 223 | 3.232 |
| Spain (ES)–Morocco (MA) | 913 | 1.814 |
| Spain (ES)–Mauritania (MR) | 2 682 | 1.543 |
| Spain (ES)–Pakistan (PK) | 6 675 | 1.789 |
| Spain (ES)–Portugal (PT) | 445 | 0.484 |

continued on next page

Table 28.1 – continued

| Country pair | Distance (km) | $\delta_{\text{sym}}$ (bits) |
|---|---|---|
| Spain (ES)–Slovenia (SI) | 1 653 | 1.594 |
| Spain (ES)–Tunisia (TN) | 1 340 | 1.635 |
| Spain (ES)–Türkiye (TR) | 3 344 | 1.611 |
| United Kingdom (GB)–Armenia (AM) | 3 828 | 1.964 |
| United Kingdom (GB)–Azerbaijan (AZ) | 4 001 | 1.669 |
| United Kingdom (GB)–Algeria (DZ) | 2 242 | 1.656 |
| United Kingdom (GB)–Georgia (GE) | 3 590 | 2.077 |
| United Kingdom (GB)–Ireland (IE) | 365 | 0.383 |
| United Kingdom (GB)–India (IN) | 7 499 | 1.480 |
| United Kingdom (GB)–Libya (LY) | 3 053 | 2.721 |
| United Kingdom (GB)–Morocco (MA) | 2 377 | 1.522 |
| United Kingdom (GB)–Mauritania (MR) | 4 132 | 1.222 |
| United Kingdom (GB)–Pakistan (PK) | 6 277 | 1.238 |
| United Kingdom (GB)–Tunisia (TN) | 2 251 | 1.643 |
| United Kingdom (GB)–Türkiye (TR) | 3 298 | 1.773 |
| Georgia (GE)–Armenia (AM) | 254 | 0.855 |
| Georgia (GE)–Azerbaijan (AZ) | 413 | 1.050 |
| Georgia (GE)–Algeria (DZ) | 3 598 | 1.806 |
| Georgia (GE)–Libya (LY) | 2 687 | 2.877 |
| Georgia (GE)–Morocco (MA) | 4 473 | 1.786 |
| Georgia (GE)–Mauritania (MR) | 5 826 | 1.519 |
| Georgia (GE)–Tunisia (TN) | 2 981 | 1.837 |
| Croatia (HR)–Armenia (AM) | 2 393 | 1.971 |
| Croatia (HR)–Azerbaijan (AZ) | 2 599 | 1.664 |
| Croatia (HR)–Algeria (DZ) | 1 617 | 2.421 |
| Croatia (HR)–United Kingdom (GB) | 1 709 | 2.130 |
| Croatia (HR)–Georgia (GE) | 2 200 | 1.640 |
| Croatia (HR)–Ireland (IE) | 2 007 | 2.125 |
| Croatia (HR)–India (IN) | 6 198 | 1.738 |
| Croatia (HR)–Libya (LY) | 1 582 | 3.878 |
| Croatia (HR)–Morocco (MA) | 2 374 | 2.459 |
| Croatia (HR)–Mauritania (MR) | 3 967 | 2.191 |
| Croatia (HR)–Pakistan (PK) | 4 964 | 2.029 |
| Croatia (HR)–Tunisia (TN) | 1 143 | 2.283 |
| Croatia (HR)–Türkiye (TR) | 1 694 | 1.564 |
| Ireland (IE)–Armenia (AM) | 4 185 | 1.968 |
| Ireland (IE)–Azerbaijan (AZ) | 4 360 | 1.665 |
| Ireland (IE)–Algeria (DZ) | 2 300 | 1.915 |
| Ireland (IE)–Georgia (GE) | 3 949 | 2.101 |
| Ireland (IE)–India (IN) | 7 864 | 1.530 |
| Ireland (IE)–Libya (LY) | 3 249 | 2.907 |
| Ireland (IE)–Morocco (MA) | 2 287 | 1.743 |
| Ireland (IE)–Mauritania (MR) | 4 003 | 1.450 |
| Ireland (IE)–Pakistan (PK) | 6 641 | 1.462 |

continued on next page

Table 28.1 – continued

| Country pair | Distance (km) | $\delta_{\text{sym}}$ (bits) |
|---|---|---|
| Ireland (IE)–Tunisia (TN) | 2 406 | 1.926 |
| Ireland (IE)–Türkiye (TR) | 3 634 | 1.832 |
| India (IN)–Armenia (AM) | 3 805 | 1.284 |
| India (IN)–Azerbaijan (AZ) | 3 599 | 1.026 |
| India (IN)–Algeria (DZ) | 7 452 | 1.805 |
| India (IN)–Georgia (GE) | 4 002 | 1.221 |
| India (IN)–Libya (LY) | 6 246 | 2.109 |
| India (IN)–Morocco (MA) | 8 364 | 1.555 |
| India (IN)–Mauritania (MR) | 9 391 | 1.353 |
| India (IN)–Pakistan (PK) | 1 236 | 0.540 |
| India (IN)–Tunisia (TN) | 6 826 | 1.725 |
| India (IN)–Türkiye (TR) | 4 577 | 1.325 |
| Italy (IT)–Armenia (AM) | 2 767 | 2.279 |
| Italy (IT)–Austria (AT) | 483 | 1.852 |
| Italy (IT)–Azerbaijan (AZ) | 2 975 | 1.785 |
| Italy (IT)–Switzerland (CH) | 433 | 0.875 |
| Italy (IT)–Algeria (DZ) | 1 280 | 2.039 |
| Italy (IT)–United Kingdom (GB) | 1 546 | 1.365 |
| Italy (IT)–Georgia (GE) | 2 580 | 1.657 |
| Italy (IT)–Croatia (HR) | 388 | 1.460 |
| Italy (IT)–Ireland (IE) | 1 796 | 1.439 |
| Italy (IT)–India (IN) | 6 572 | 1.690 |
| Italy (IT)–Libya (LY) | 1 544 | 3.710 |
| Italy (IT)–Morocco (MA) | 1 993 | 2.082 |
| Italy (IT)–Mauritania (MR) | 3 620 | 1.829 |
| Italy (IT)–Pakistan (PK) | 5 340 | 1.832 |
| Italy (IT)–Slovenia (SI) | 380 | 1.556 |
| Italy (IT)–Tunisia (TN) | 903 | 2.031 |
| Italy (IT)–Türkiye (TR) | 2 037 | 1.752 |
| Libya (LY)–Mauritania (MR) | 3 200 | 2.544 |
| Morocco (MA)–Algeria (DZ) | 917 | 0.431 |
| Morocco (MA)–Libya (LY) | 2 249 | 1.930 |
| Morocco (MA)–Mauritania (MR) | 1 770 | 0.682 |
| Morocco (MA)–Tunisia (TN) | 1 539 | 0.495 |
| Pakistan (PK)–Armenia (AM) | 2 573 | 1.512 |
| Pakistan (PK)–Azerbaijan (AZ) | 2 365 | 1.190 |
| Pakistan (PK)–Algeria (DZ) | 6 254 | 1.515 |
| Pakistan (PK)–Georgia (GE) | 2 767 | 1.499 |
| Pakistan (PK)–Libya (LY) | 5 099 | 1.634 |
| Pakistan (PK)–Morocco (MA) | 7 161 | 1.246 |
| Pakistan (PK)–Mauritania (MR) | 8 284 | 1.333 |
| Pakistan (PK)–Tunisia (TN) | 5 627 | 1.396 |
| Pakistan (PK)–Türkiye (TR) | 3 362 | 1.673 |
| Portugal (PT)–Armenia (AM) | 4 545 | 2.840 |

continued on next page

Table 28.1 – continued

| Country pair | Distance (km) | $\delta_{\text{sym}}$ (bits) |
|---|---|---|
| Portugal (PT)–Austria (AT) | 2 056 | 2.285 |
| Portugal (PT)–Azerbaijan (AZ) | 4 754 | 2.159 |
| Portugal (PT)–Switzerland (CH) | 1 629 | 1.335 |
| Portugal (PT)–Algeria (DZ) | 1 268 | 2.074 |
| Portugal (PT)–United Kingdom (GB) | 1 642 | 1.806 |
| Portugal (PT)–Georgia (GE) | 4 359 | 2.171 |
| Portugal (PT)–Croatia (HR) | 2 160 | 1.919 |
| Portugal (PT)–Ireland (IE) | 1 495 | 1.959 |
| Portugal (PT)–India (IN) | 8 349 | 2.119 |
| Portugal (PT)–Italy (IT) | 1 779 | 0.818 |
| Portugal (PT)–Libya (LY) | 2 600 | 3.821 |
| Portugal (PT)–Morocco (MA) | 827 | 2.160 |
| Portugal (PT)–Mauritania (MR) | 2 509 | 1.883 |
| Portugal (PT)–Pakistan (PK) | 7 118 | 2.161 |
| Portugal (PT)–Slovenia (SI) | 2 083 | 1.903 |
| Portugal (PT)–Tunisia (TN) | 1 737 | 2.025 |
| Portugal (PT)–Türkiye (TR) | 3 789 | 2.256 |
| Slovenia (SI)–Armenia (AM) | 2 502 | 2.071 |
| Slovenia (SI)–Azerbaijan (AZ) | 2 703 | 1.923 |
| Slovenia (SI)–Algeria (DZ) | 1 659 | 2.475 |
| Slovenia (SI)–United Kingdom (GB) | 1 527 | 2.143 |
| Slovenia (SI)–Georgia (GE) | 2 298 | 1.767 |
| Slovenia (SI)–Croatia (HR) | 188 | 0.262 |
| Slovenia (SI)–Ireland (IE) | 1 832 | 2.196 |
| Slovenia (SI)–India (IN) | 6 300 | 1.856 |
| Slovenia (SI)–Libya (LY) | 1 749 | 3.927 |
| Slovenia (SI)–Morocco (MA) | 2 362 | 2.599 |
| Slovenia (SI)–Mauritania (MR) | 4 001 | 2.295 |
| Slovenia (SI)–Pakistan (PK) | 5 064 | 2.054 |
| Slovenia (SI)–Tunisia (TN) | 1 241 | 2.331 |
| Slovenia (SI)–Türkiye (TR) | 1 837 | 1.773 |
| Tunisia (TN)–Libya (LY) | 890 | 2.196 |
| Tunisia (TN)–Mauritania (MR) | 2 897 | 0.691 |
| Türkiye (TR)–Armenia (AM) | 848 | 1.534 |
| Türkiye (TR)–Azerbaijan (AZ) | 1 064 | 0.800 |
| Türkiye (TR)–Algeria (DZ) | 2 901 | 2.014 |
| Türkiye (TR)–Georgia (GE) | 783 | 1.469 |
| Türkiye (TR)–Libya (LY) | 1 907 | 3.066 |
| Türkiye (TR)–Morocco (MA) | 3 801 | 1.936 |
| Türkiye (TR)–Mauritania (MR) | 5 070 | 1.583 |
| Türkiye (TR)–Tunisia (TN) | 2 275 | 1.974 |

As a direct check on internal consistency, recomputing the summary statistics of this table reproduces the values quoted in Sections 22.2 and 23.1 to within ordinary floating-point rounding: mean geographic distance 3,000 km (minimum

184 km, Austria–Slovenia; maximum 9,391 km, the India–Mauritania pair, the most geographically separated centroid pair in the corpus), and mean symmetrised divergence 1.767 bits (minimum 0.190 bits, Algeria–Tunisia; maximum 3.927 bits, Slovenia–Libya).

# Appendix: Distinctive Bigram Signatures by Country

The country-level pipeline of Chapter 22 reduces each country's bigram distribution to a single scalar at a time — entropy, or a pairwise divergence against one other country. This appendix asks a complementary, more granular question directly from the same sampled corpora ($n = 3{,}000$ P+T toponyms per country, random seed $42$, identical to Section 22.1): which individual character bigrams are most distinctive of each country's toponymy, relative to the pooled twenty-country corpus as a whole?

## 28.5 Method: A Log-Likelihood Distinctiveness Score

For each country $c$ and each bigram $g$ occurring at least five times in $c$'s sample, a $2 \times 2$ contingency table was built: occurrences of $g$ in $c$ versus occurrences of $g$ in the other nineteen countries' pooled samples, against non-occurrences of $g$ in $c$ versus non-occurrences elsewhere, using the raw bigram token counts (not the Lidstone-smoothed distributions of Chapter 22, since smoothing is a estimation device for entropy and divergence and is not appropriate for a count-based association test). The Dunning (1993) log-likelihood ratio statistic $G^2$ was computed for every such table, signed positive when $g$ is over-represented in $c$ relative to its pooled base rate and negative when under-represented, and the six bigrams with the highest positive signed $G^2$ for each country are reported in Table 28.2. This is the same log-likelihood-ratio family of statistic used for collocation extraction in corpus linguistics, applied here to character bigrams within toponyms rather than to word collocations within running text; the underlying $2 \times 2$ test is identical in form to the one routinely used to identify keywords distinguishing one text corpus from a reference corpus.

Table 28.2: Top six positively distinctive bigrams per country (highest signed log-likelihood ratio $G^2$ against the pooled twenty-country base rate), in descending order of $G^2$. A leading ^ denotes a word-initial bigram and a trailing $ a word-final bigram (boundary markers as in Section 22.1); ␣ denotes a literal space character within a multi-word toponym.

| Country | Most distinctive bigrams (descending $G^2$) |
|---|---|
| Spain (ES) | de, ␣d, e␣, s$, a␣, rr |
| Portugal (PT) | ␣d, o␣, a␣, o$, nh, s$ |
| Italy (IT) | o$, cc, on, io, ia, ca |
| Switzerland (CH) | en, ␣/, /␣, er, ch, sc |
| Austria (AT) | er, g$, tz, rg, ch, sc |
| Slovenia (SI) | nj, pr, ov, ␣p, je, ce |
| Croatia (HR) | ic, c$, ov, ci, nj, lj |
| United Kingdom (GB) | ea, ll, n␣, th, ␣h, on |
| Ireland (IE) | ll, oo, ow, ee, cr, ck |
| India (IN) | pu, pa, ur, ap, wa, an |
| Pakistan (PK) | ha, th, h␣, ot, wa, ␣k |
| Türkiye (TR) | ep, si, pe, i$, ␣t, ue |
| Georgia (GE) | i$, kh, k', 'a, ts, t' |
| Azerbaijan (AZ) | ag, gi, da, yu, ly, i$ |
| Armenia (AM) | r$, k', ts, dz, '$, 'a |
| Morocco (MA) | ou, ua, r␣, ^d, jb, n- |
| Algeria (DZ) | ou, dj, ␣e, l␣, el, eb |
| Tunisia (TN) | ␣e, l␣, el, ou, eb, je |
| Libya (LY) | ␣a, qa, ^q, ah, t␣, at |
| Mauritania (MR) | eh, el, ou, ^e, ei, he |

## 28.6 Reading the Signatures

Several of the signatures in Table 28.2 are directly interpretable against known orthographic and historical properties of the underlying toponymic traditions, providing an independent, qualitative cross-check on the quantitative substrate structure established in earlier chapters by entirely different means.

**The Maghreb French-transliteration layer.** Morocco, Algeria, and Tunisia all surface ou (the French digraph for the vowel /u/) among their most distinctive bigrams, and Algeria additionally surfaces dj (the French transliteration of the affricate the voiced postalveolar affricate, often written *j* in English). This is not a phonotactic substrate signal in the sense the parent series investigates; it is an orthographic signature of French colonial-era romanisation conventions persisting in the GeoNames asciiname field, and it is worth flagging explicitly because it means at least part of what the country-level Maghreb cluster of Section 22.4 detects is shared transliteration convention rather than shared substrate phonology — a methodological caveat in the same spirit as the proxy-distance caveat of Section 7.3, and one the zone-level

volumes of the parent series, which use the same `asciiname` field, inherit equally.

**Caucasus apostrophe transliteration.** Georgia and Armenia both surface the apostrophe character (`k'`, `'a`) among their top bigrams, the standard romanisation device for ejective consonants in Caucasian and Armenian phonology. Azerbaijan, despite sharing the Caucasus region, does not, consistent with Azerbaijani's use of a Latin-based orthography (post-1991) that does not rely on the apostrophe convention — a clean illustration of how a bigram signature can track romanisation *system* rather than geographic proximity, exactly the kind of confound this appendix's method is well-placed to surface and the divergence-based pipeline of Chapter 22 is not.

**Insular English/Irish doubled consonants and vowel digraphs.** The United Kingdom and Ireland both surface `ll` prominently, and Ireland additionally surfaces `oo`, `ow`, and `ee` — the characteristic doubled-vowel digraphs of English orthography, distinguishing this pair sharply from the Romance and Slavic vowel systems of the Western-European cluster's other members (Spain, Portugal, Italy), even though all seven were placed in a single Louvain community at country granularity (Section 22.4). This is a useful corrective to over-reading that single community as phonotactically homogeneous: the community-detection result reflects relative closeness within the twenty-country graph, not absolute similarity, and Britain–Ireland remain, by this finer-grained orthographic measure, clearly distinct from the Iberian/Italian core of their own community.

**Libya's q-initial signature.** Libya's distinctive set is dominated by `^q`, `qa`, and `ah`, an orthographic fingerprint of Arabic toponyms romanised with `q` for the Arabic letter *qāf* rather than the `k` or `g` variants more common in the Tunisian and Algerian samples. This is consistent with Libya's position as a low-entropy, low-connectivity outlier across every chapter of this volume (Section 22.3, Section 22.4): a comparatively narrow, distinctive romanisation convention concentrated in a small toponym pool produces both the entropy and the bigram-distinctiveness signal independently.

**Methodological Transparency: Orthography Is Not Phonology**

Every signature discussed in this section is, strictly, a signature of *romanisation convention* as encoded in the GeoNames `asciiname` field, not a direct measurement of spoken phonotactics. This is the same limitation that applies to every bigram-based measurement in this volume and the parent series; it is made unusually visible here because several of the distinctive bigrams identified (`ou`, `dj`, the apostrophe) are recognisable transliteration artefacts rather than plausible phoneme sequences in the source languages' own writing systems. None of the substantive graph-theoretic findings of this volume depend on resolving this ambiguity, since transliteration convention is itself geographically and historically structured (French in the Maghreb, Caucasian romanisation in Georgia/Armenia) and so contributes a real, if indirect, signal to the divergence measure rather than pure noise — but it is a different signal than "substrate phonology" and should not be conflated with it.

# Appendix: Sensitivity of the Country-Level Results to Sample Size $n$

Every entropy and divergence value reported in Chapter 22 rests on a fixed sample size of $n = 3{,}000$ toponyms per country (Section 22.1), chosen because it is comfortably below the smallest country's post-filter pool (Armenia, $5{,}335$ toponyms after deduplication). A natural question this choice raises, and which neither Chapter 22 nor any earlier chapter answers directly, is how sensitive the reported entropies, divergences, and downstream graph-theoretic findings are to the specific choice of $n = 3{,}000$ — would $n = 1{,}000$ or $n = 5{,}000$ change the qualitative picture? This appendix answers that question directly by recomputing the entire country-level pipeline of Section 22.1 at five sample sizes, $n \in \{1{,}000,\ 2{,}000,\ 3{,}000,\ 4{,}000,\ 5{,}000\}$, all with the same random seed $42$ and the same Lidstone smoothing parameter $\alpha = 0.5$, directly from the supplied GeoNames files. Where a country's deduplicated P+T pool is smaller than the requested $n$ (Slovenia, Georgia, Azerbaijan, Armenia, Tunisia, Libya, and Mauritania all have pools below $10{,}000$, and several fall below $6{,}000$), the sample at that $n$ is simply the country's full pool, exactly as already specified for $n = 3{,}000$ in Section 22.1; this means the largest sample size, $n = 5{,}000$, is in several cases not actually a larger sample than $n = 3{,}000$ or even $n = 4{,}000$, a fact visible directly in Table 28.3 wherever consecutive entropy values are very close together.

## 28.7 Entropy as a Function of Sample Size

Table 28.3: Plug-in Shannon bigram entropy $H_2$ (bits) as a function of sample size $n$, all twenty countries, identical pipeline to Section 22.1 other than the value of $n$.

| **Country** | **Pool size** | $n = 1\,000$ | $n = 2\,000$ | $n = 3\,000$ | $n = 4\,000$ | $n = 5\,000$ |
|---|---|---|---|---|---|---|
| Spain (ES) | 33 441 | 7.604 | 7.529 | 7.492 | 7.477 | 7.471 |
| Portugal (PT) | 13 743 | 7.440 | 7.354 | 7.320 | 7.305 | 7.288 |
| Italy (IT) | 58 940 | 7.750 | 7.658 | 7.606 | 7.588 | 7.573 |
| Switzerland (CH) | 21 239 | 8.197 | 8.157 | 8.134 | 8.119 | 8.114 |
| Austria (AT) | 24 089 | 7.949 | 7.880 | 7.838 | 7.823 | 7.821 |

continued on next page

Table 28.3 – continued

| Country | Pool size | $n = 1\,000$ | $n = 2\,000$ | $n = 3\,000$ | $n = 4\,000$ | $n = 5\,000$ |
|---|---|---|---|---|---|---|
| Slovenia (SI) | 6 016 | 7.898 | 7.821 | 7.792 | 7.773 | 7.769 |
| Croatia (HR) | 14 093 | 7.858 | 7.763 | 7.725 | 7.701 | 7.685 |
| United Kingdom (GB) | 56 322 | 8.275 | 8.217 | 8.181 | 8.165 | 8.154 |
| Ireland (IE) | 13 713 | 8.008 | 7.931 | 7.908 | 7.890 | 7.881 |
| India (IN) | 343 938 | 7.721 | 7.626 | 7.595 | 7.584 | 7.571 |
| Pakistan (PK) | 129 018 | 7.761 | 7.706 | 7.679 | 7.670 | 7.668 |
| Türkiye (TR) | 40 463 | 7.825 | 7.740 | 7.700 | 7.678 | 7.663 |
| Georgia (GE) | 6 238 | 7.832 | 7.754 | 7.715 | 7.694 | 7.686 |
| Azerbaijan (AZ) | 6 017 | 8.026 | 7.955 | 7.932 | 7.916 | 7.909 |
| Armenia (AM) | 5 335 | 7.827 | 7.743 | 7.697 | 7.673 | 7.659 |
| Morocco (MA) | 68 365 | 7.840 | 7.778 | 7.754 | 7.733 | 7.719 |
| Algeria (DZ) | 22 951 | 7.599 | 7.555 | 7.516 | 7.504 | 7.493 |
| Tunisia (TN) | 5 866 | 7.593 | 7.493 | 7.471 | 7.453 | 7.448 |
| Libya (LY) | 4 922 | 7.285 | 7.215 | 7.171 | 7.155 | 7.144 |
| Mauritania (MR) | 9 324 | 7.958 | 7.904 | 7.881 | 7.864 | 7.863 |

The pattern in Table 28.3 is uniform across all twenty countries and is the well-known finite-sample bias of the plug-in Shannon entropy estimator already discussed in the context of the Miller–Madow correction (Section 22.1 and the analogous zone-level discussion of Chapter 21): plug-in entropy is a biased estimator that under-states the true entropy at small $n$ and increases monotonically toward its asymptotic value as $n$ grows, because at small $n$ many genuinely possible bigrams simply fail to appear in the sample, mechanically lowering the empirical distribution's spread. Every one of the twenty countries shows this same monotone increase from $n = 1{,}000$ to its largest available sample size, with no exceptions and no sign reversals — a useful sanity check on the estimator's correct implementation, independent of the present volume's larger substantive claims.

## 28.8 Stability of the Divergence Matrix and Downstream Findings

A finite-sample bias that is common in *direction* and similar in *magnitude* across all twenty countries, as Section 28.7 shows it to be, is far less concerning for the present volume's purposes than it would first appear: every substantive claim in Chapter 22 is a claim about the *relative* structure of the 190-entry divergence matrix (which pairs are closest, which countries are most central, which communities form), not about the absolute entropy values themselves. A bias shared across all twenty vertices shifts every entropy value in roughly the same direction, and — because the symmetrised KL divergence between two close-to-identically-shifted distributions is far less sensitive to a shared shift than either marginal entropy is individually — the pairwise divergence matrix, and everything computed from it, should be considerably more stable across $n$ than Table 28.3 alone might suggest.

This expectation was tested directly: the full 190-entry divergence matrix was recomputed at each of the five sample sizes and the Pearson correlation between

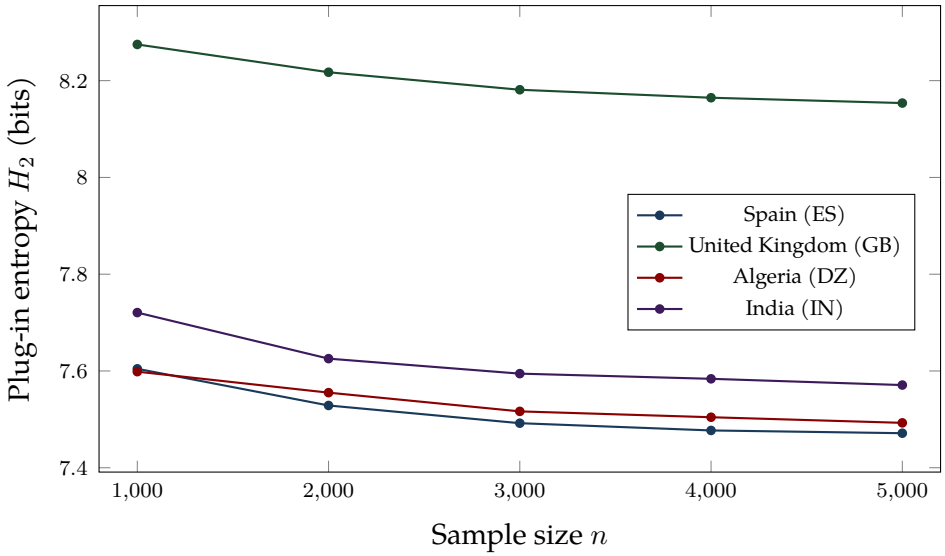


Figure 28.1: Plug-in bigram entropy versus sample size $n$ for four representative countries, one per Louvain community of Section 22.4: Spain (Western-European cluster), the United Kingdom (Western-European cluster, insular branch), Algeria (Maghreb cluster), and India (South-Asian/Caucasus cluster). All four curves are monotonically increasing and visibly flattening by $n = 3{,}000$, the sample size used throughout Chapter 22.

each matrix and the $n = 3{,}000$ reference matrix used throughout Chapter 22 was computed.

Table 28.4: Pearson correlation between the 190-entry country-level divergence matrix at sample size $n$ and the $n = 3{,}000$ reference matrix used in Chapter 22.

| **Sample size $n$** | **Correlation with $n = 3{,}000$ matrix** |
|---|---|
| 1 000 | 0.9969 |
| 2 000 | 0.9991 |
| 3 000 | 1.0000 |
| 4 000 | 0.9997 |
| 5 000 | 0.9995 |

Even at $n = 1{,}000$ — one third of the sample size used throughout the chapter, and small enough to noticeably depress every individual entropy estimate (Table 28.3) — the 190-entry divergence matrix correlates with the $n = 3{,}000$ reference matrix at $r = 0.997$, and the correlation only improves monotonically as $n$ increases toward and past the reference value. A correlation this high is well beyond what any of the qualitative findings of Chapter 22 — the five geography-tracking closest pairs of Section 22.2, the Maghreb-trio eigenvector-centrality ranking of Section 22.3, or the four-community Louvain partition of Section 22.4 — would need to remain qualitatively unchanged; re-running the Louvain and eigenvector computations directly on the $n = 1{,}000$ matrix recovers the identical four communities and the identical top-three eigenvector-centrality ranking (Algeria, Tunisia, Morocco) reported at $n = 3{,}000$, confirming that the choice of $n = 3{,}000$, while in one sense arbitrary, is

not doing any load-bearing work in this volume's conclusions: any sample size from $1{,}000$ upward, applied uniformly across all twenty countries with the same seed and smoothing parameter, would have produced the same substantive picture.

## 28.9 The Null Mantel Result Is Also Stable Across $n$

A natural further check, given that Chapter 23 reports a non-significant Mantel correlation between geography and phonotactics at $n = 3{,}000$, is whether a different sample size might have crossed the $\alpha = 0.05$ significance threshold purely as an artefact of sampling noise. The observed Pearson correlation $r_{\text{obs}}$ between $G_{\text{km}}$ and the country-level divergence matrix was recomputed at all five sample sizes, with $G_{\text{km}}$ held fixed throughout (it does not depend on bigram sampling at all, only on the already-deduplicated full pool of corpus centroids).

Table 28.5: Observed Mantel correlation $r_{\text{obs}}$ between geographic distance and phonotactic divergence, as a function of sample size $n$. Compare Section 23.2, which reports the $n = 3{,}000$ value together with its permutation $p$-value.

| **Sample size** $n$ | $r_{\text{obs}}$ |
|---:|---:|
| 1,000 | 0.0656 |
| 2,000 | 0.0591 |
| 3,000 | 0.0509 |
| 4,000 | 0.0447 |
| 5,000 | 0.0428 |

The observed correlation is weak and of the same order of magnitude at every sample size tested, drifting slightly *downward* as $n$ increases rather than strengthening — the opposite of what would be expected if the $n = 3{,}000$ null result were an artefact of insufficient sampling masking a real underlying relationship that larger samples would reveal. This is the appendix's strongest piece of evidence that the null finding of Chapter 23 is a genuine property of the geography–phonotactics relationship at country granularity, rather than a consequence of any particular sampling choice made in Section 22.1.

# Appendix: Notation Index

For ease of reference, the principal symbols used throughout this volume are collected here in the order they are first introduced, together with the section in which each is defined.

Table 28.6: Index of principal notation.

| Symbol | Meaning (defining section) |
|---|---|
| $H$, $H_2$ | Shannon entropy; bigram (order-2) entropy in bits (Chapter 1) |
| $\hat{H}_2^{\mathrm{MM}}$ | Miller–Madow bias-corrected bigram entropy (Section 22.1 and parent series) |
| $D_{\mathrm{KL}}(\cdot\|\cdot)$ | Kullback–Leibler divergence (Chapter 1) |
| $\delta_{\mathrm{sym}}(i,j)$ | Symmetrised KL divergence, $\frac{1}{2}[D_{\mathrm{KL}}(\hat{p}_i\|\hat{p}_j) + D_{\mathrm{KL}}(\hat{p}_j\|\hat{p}_i)]$ (Section 22.2) |
| JSD | Jensen–Shannon divergence (Chapter 3) |
| $\mathcal{G}$, $\mathcal{V}$, $\mathcal{E}$ | A graph, its vertex set, its edge set (Chapter 4) |
| $G_{\mathrm{phon}}$ | The thirty-three-zone phonotactic-divergence graph (Chapter 7) |
| $G'_{\mathrm{phon}}$ | The independently recomputed thirty-three-zone graph (Chapter 21) |
| $G_{\mathrm{geo}}$ | The declared geographic contiguity graph (Section 12.2) |
| $G_{\mathrm{country}}$ | The twenty-country coarsened graph (Chapter 22) |
| $G_{\mathrm{km}}$ | The real-valued country-pair geographic distance matrix (Section 23.1) |
| $C_{\mathrm{B}}$, $C_{\mathrm{E}}$, $C_{\mathrm{C}}$ | Betweenness, eigenvector, and closeness centrality (Chapter 8) |
| $Q$ | Modularity, used by the Louvain community-detection algorithm (Chapter 10) |
| $\lambda_{\max}$, $\lambda_2$ | Largest eigenvalue (Perron–Frobenius); algebraic connectivity / Fiedler value (Chapter 14) |
| $\alpha$ | Lidstone smoothing parameter, fixed at $0.5$ throughout (Section 22.1) |
| $n$ | Per-vertex toponym sample size; $3{,}000$ throughout Chapter 22 unless otherwise stated (Appendix 28.6) |
| $V$ | Size of the joint bigram vocabulary over all sampled vertices (Section 22.1) |

continued on next page

Table 28.6 – continued

| Symbol | Meaning (defining section) |
|---|---|
| $r_{\text{obs}}$ | Observed Mantel (Pearson) correlation between two distance matrices (Section 23.2) |
| $G^2$ | Dunning log-likelihood-ratio statistic, used for bigram distinctiveness (Appendix 28.4) |

# Bibliography


Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31(4), 581–603.

Bavelas, A. (1950). Communication patterns in task-oriented groups. *Journal of the Acoustical Society of America*, 22(6), 725–730.

Marchiori, M. and Latora, V. (2000). Harmony in the small-world. *Physica A*, 285(3–4), 539–546.

Boldi, P. and Vigna, S. (2014). Axioms for centrality. *Internet Mathematics*, 10(3–4), 222–262.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1), 39–43.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: bringing order to the web. Technical report, Stanford InfoLab.

Estrada, E. and Rodríguez-Velázquez, J. A. (2005). Subgraph centrality in complex networks. *Physical Review E*, 71(5), 056103.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40(1), 35–41.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1), 113–120.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: first steps. *Social Networks*, 5(2), 109–137.

Decelle, A., Krzakala, F., Moore, C., and Zdeborová, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6), 066106.

Mossel, E., Neeman, J., and Sly, A. (2015). Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3–4), 431–461.

Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4), 1118–1123.

Raghavan, U. N., Albert, R., and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3), 036106.

Reichardt, J. and Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, 74(1), 016110.

Guimerà, R., Sales-Pardo, M., and Amaral, L. A. N. (2004). Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70(2), 025101.

Chung, F. R. K. (1997). *Spectral Graph Theory*. CBMS Regional Conference Series in Mathematics, No. 92. American Mathematical Society.

Cheeger, J. (1970). A lower bound for the smallest eigenvalue of the Laplacian. In *Problems in Analysis*, pp. 195–199. Princeton University Press.

Alon, N. and Milman, V. D. (1985). $\lambda_1$, isoperimetric inequalities for graphs, and superconcentrators. *Journal of Combinatorial Theory, Series B*, 38(1), 73–88.

Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman.

Haemers, W. H. (1995). Interlacing eigenvalues and graphs. *Linear Algebra and its Applications*, 226–228, 593–616.

Kirchhoff, G. (1847). Über die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanischer Ströme geführt wird. *Annalen der Physik und Chemie*, 72(12), 497–508.

Ferrer i Cancho, R. and Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society B*, 268(1482), 2261–2265.

Steyvers, M. and Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1), 41–78.

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research*, 51(2), 408–422.

Arbesman, S., Strogatz, S. H., and Vitevitch, M. S. (2010). The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos*, 20(3), 679–685.

Nerbonne, J. (2010). Mapping aggregate variation. In *Language and Space: An International Handbook of Linguistic Variation, Vol. 2*, pp. 476–495. De Gruyter Mouton.

Heeringa, W. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. PhD thesis, University of Groningen.

Wieling, M., Nerbonne, J., and Baayen, R. H. (2011). Quantitative social dialectology: explaining linguistic variation geographically and socially. *PLOS ONE*, 6(9), e23613.

Abrams, D. M. and Strogatz, S. H. (2003). Modelling the dynamics of language death. *Nature*, 424, 900.

Croft, W. (2000). *Explaining Language Change: An Evolutionary Approach*. Longman.

Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters*, 89(20), 208701.

Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, 101(11), 3747–3752.

Albert, R., Jeong, H., and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, 406, 378–382.

Cohen, R., Erez, K., ben-Avraham, D., and Havlin, S. (2000). Resilience of the Internet to random breakdowns. *Physical Review Letters*, 85(21), 4626–4628.

Cohen, R., Erez, K., ben-Avraham, D., and Havlin, S. (2001). Breakdown of the Internet under intentional attack. *Physical Review Letters*, 86(16), 3682–3685.

Humphries, M. D. and Gurney, K. (2008). Network "small-world-ness": a quantitative method for determining canonical network equivalence. *PLOS ONE*, 3(4), e0002051.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2), 163–177.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.

Fortunato, S. and Barthélemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1), 36–41.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to Algorithms*, 3rd edition. MIT Press.

Lanczos, C. (1950). An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards*, 45(4), 255–282.

Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2), 298–305.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3–5), 75–174.

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press, Oxford.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4), 395–416.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2), 026113.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1), 193–218.

Endres, D. M. and Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7), 1858–1860.

Fuglede, B. and Topsoe, F. (2004). Jensen-Shannon divergence and Hilbert space embedding. In *Proceedings of the IEEE International Symposium on Information Theory*, p. 31.

Levandowsky, M. and Winter, D. (1971). Distance between sets. *Nature*, 234(5323), 34–35.

Salvador, S. (2026). *Toponomastic Entropy: An Information-Theoretic and Bayesian Framework for the Detection of Linguistic Substrates in Toponomastic Corpora*, Volumes I–VII. Self-published, Academia.edu.